BAYESIAN NETWORKS
Bayesian reasoning in court

Anna Latour, s1456571
a.l.d.latour@umail.leidenuniv.nl
Sunday 21st June, 2015

Abstract

This work provides an overview of Bayesian reasoning in court. We give a general framework for reasoning with uncertainty in court. We address different probabilistic fallacies that people commit when reasoning with uncertainty. We also discuss different heuristics people use when assessing subjective probabilities, and the biases that these heuristic might cause. We discuss different design choices that can be made when creating a Bayesian model, and different ways of interpreting Bayesian reasoning. Finally, we discuss a proposal to use Bayesian Networks to visualise abstract mathematical models in court, in order to help the jury avoid fallacies as described in this work. We find that the benefits of this approach are possibly limited, although we do acknowledge that it might help to make a jury more aware of how Bayesian reasoning works.

1 Introduction

What is the most blatant example of the unjust conviction of an innocent person in recent history that you can call to mind? What was the evidence against this person? What cracks appeared in this evidence that finally brought to light the person's innocence? Chances are some of the evidence against this person was based on statistics. Wrong statistics.

In the Netherlands, a striking example is that of Lucia de Berk, a nurse who was sentenced life-time imprisonment in 2003, being held responsible for four murders and three attempted murders. Part of the evidence that got her convicted, consisted of the notion that a surprisingly large number of people died during Lucia's shifts at the hospital, and that their deaths could not be explained by the doctors. Chances of this happening were estimated by forensic psychologist Henk Elffers to be one in 342 million. Professor of Statistics Richard Gill reckons that this number is what decided Lucia's faith [6]. Judges, journalists, members of the public all had this number in the back of their minds, probably stopping to question her guilt.

Later, after philosopher Ton Derksen, doctor Metta de Noo, and professors of Statistiscs Philip Dawid, Richard Gill and Piet Groeneboom had worked [2, 7, 12] to draw the public's attention to the ways in which wrong statistics were (mis)used in the conviction of Lucia, the case was reopened. In 2011, Lucia de Berk was acquitted, the court judging that all people she was said to have murdered, actually died natural deaths.

In this work, we address several questions related to incidents like this. What is the role of statistics and probability in court cases? Why do we need to reason with uncertainty in court? Who is responsible for computing and estimating probabilities? What mistakes can occur while computing or estimating probabilities? How should probabilistic pieces of evidence be combined to yield a judgement about somebody's guilt? How does the way that probabilistic evidence is presented to court influence justice?

This work is organised as follows. We present a framework for reasoning with uncertainty in court in Section 2, and use that in the rest of the work to explain certain difficulties with reasoning with uncertainty. We address common biases that are found when both laypeople and experts make subjective probabilistic estimates in Section 4. We discuss common probabilistic fallacies made when reasoning with uncertainty in Section 3. Section 5 highlights some of the different ways of looking at Bayesian reasoning, and of constructing Bayesian models. In Section 6, we highlight different approaches to presenting reasoning with uncertainty (and Bayesian reasoning in particular) in courts. Finally, we conclude this work in Section 7.

2 The basics

In this section we provide a theoretical framework for this work. We describe the basics of reasoning with uncertainty in court.

2.1 A simple problem

Consider the simplest form of reasoning with uncertainty that one might come across during a court case. A piece of evidence is presentented, and the judge or jury has to decide if this piece of evidence serves to increase, decrease or not alter their belief in the defendant's guilt. Suppose, for example, that a sample of DNA was found at the crime scene. This fact can be seen in the context of the causal chain of events presented in Figure 1.

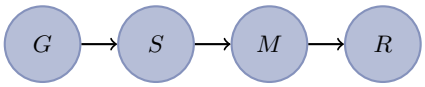

Figure 1: A typical causal chain of events that might be considered in a court case. If the defendant is guilty of committing a crime (G), he or she may be the source (S) some evidence that was left at or near the crime scene, and that can be directly linked to him or her. If a sample is taken from the defendant, it might match the aforementioned evidence (M). Finally, when investigating the match, a lab may report that the sample from the defendant does indeed match the evidence found at the crime scene (R).

Here, suppose that a person is *guilty* of committing a

certain crime. Suppose that a piece of evidence is found at the crime scene, for example some saliva containing DNA, or CCTV footage showing a car at the crime scene at the moment that the crime took place. This evidence can to some extent be *linked* to the defendant. DNA or fingerprints can be very directly linked to a specific person, but it might also be that the defendant owns a car of the same make, model and colour as shown in the CCTV footage. Now, there is a possibility that the evidence found at the crime scene does indeed *match* a sample taken from the defendant. For example, the DNA in the saliva does belong to the defendant, or it is his/her car that is on the CCTV footage. Finally, a test may report that a match is found between the sample from the defendant and the sample from the crime scene. For example, a DNA match is found, or the make, model and colour of the defendant's car can be recognised from the CCTV.

This only concerns one piece of evidence. In a real court case, a judge or jury may be confronted with different pieces of evidence, all being of different types (DNA, footprint, fingerprint, CCTV footage, audio, etcetera), and all of them may or may not be dependent of one another.

Even correctly interpreting and using the very basic causal chain of events described above, represents a true challenge for anyone confronted with it in court. Firstly, there are several probabilities that have to be estimated, which in some cases can only be done so *subjectively*. This means that the person making these estimation does not only have to be an expert in the relevant field, but has to also beware of being *biased* towards certain values. Secondly, although the chain is simple, it is easy to confuse certain probabilities with others, and thus to commit *probabilistic fallacies*, especially when being questioned by an attorney.

2.2 Likelihood ratio and posterior odds

The simple chain of reasoning described above is not sufficient for making decision in court. Typically, pieces of evidence should be presented in the context of a *Likelyhood Ratio* (LR) that describes the ratio between the probabilities of getting a result under two different hypotheses. For example, suppose that a footprint is found at the crime scene, and a forensic expert is comparing this footprint with the shoe of the defendant. The expert will consider two hypotheses:

H_1**:** the shoe of the defendant is the source of the footprint at the crime scene;

H_2**:** another shoe with a similar profile and size is the source of the footprint at the crime scene.

Note that these hypotheses are mutually exclusive and exhaustive. The similarities and differences between the footprint and the defendant's shoe represent the result of the expert's investigation. If they fit H_1 better than they fit H_2, this forms evidence for H_1. This might happen if there is, for example, a little stone caught in the profile of the defendant's shoe, and the footprint shows a little mark of the same size at exactly the same spot. On the other hand, when there are no striking features that the expert can find, the evidence for H_1 becomes weaker, and both hypotheses might be about equally likely. Similarly, the defendant's shoe might show saliencies that are not present in the footprint, which might make H_2 more likely.

These findings are summarised in the LR, which is defined as

$$LR = \frac{P\left(\text{results} \mid H_1\right)}{P\left(\text{results} \mid H_2\right)}. \tag{1}$$

A different way of using probability in court is by using the ratio of the *posterior odds* of H_1 being true and H_2 being true. These odds are computed as follows:

$$\frac{P\left(H_1 \mid \text{results}\right)}{P\left(H_2 \mid \text{results}\right)} = \frac{P\left(\text{results} \mid H_1\right)}{P\left(\text{results} \mid H_2\right)} \times \frac{P\left(H_1\right)}{P\left(H_2\right)}, \tag{2}$$

where the first term on the right hand side represents the LR, and the second term represents the *prior odds*, or the ratio of probabilities of H_1 being true and H_2 being true before any evidence is taken into consideration. The posterior odds represent the ratio between these probabilities if the results of the investigation of the piece of evidence are taken into account. Note that this expression follows directly from Bayes' rule, and eliminates the need to assess prior probability $P(\text{results})$.

3 Fallacies in reasoning with uncertainty

In this section, we discuss a range of different fallacies that may be committed by people working with probabilities. We discuss two main types: fallacies caused by a lack of understanding of conditional probability in dealing with single pieces of evidence, and fallacies caused by a lack of understanding on how to combine different pieces of evidence.

3.1 Probabilistic fallacies

Consider the simple causal chain as presented in Figure 1. In this section, we explain a number of common fallacies, referring to the aforementioned chain. We indicate the statement that the defendant is guilty of committing the crime with g, and its negation with $\overline{g}$. Similarly, the statement that the defendant is the source of the evidence found, is indicated with s, and its negation with $\overline{s}$, etcetera. This section is based on the work presented in [10, 5]

Let us start at the end of our example chain of reasoning, with the probability $P(r)$. It is noted in [10] that DNA analysts tend to overestimate the reliability of DNA analysis. They tend to claim that the method is failsafe, and thus yields no false positives. Thus, they assume that $P\left(r \mid m\right) = 1$, or even $P\left(m \mid s\right) = 1$ or $P\left(s \mid g\right) = 1$! This is known as the *Impossibility of False Positives* fallacy. Koehler also notes in [10] that his analysis of a report on by the Collaborative Testing Services yields an estimate of false positive errors occurring in one to four percent of match reports provided by labs in *open* proficiency tests. Having no data on false positive rates in *closed* proficiency tests, staged to mimic realistic crime scene conditions, it is probably not unreasonable to assume a false positive rate in these conditions to be of the same order of magnitude.

The *Interrogator's Fallacy* occurs when the evidence consists of a confession of guilt from the defendant (that is not corroborated). Here, $P(r \mid g)$ is used to infer $P(g \mid r)$ without taking into account the probabil-

ity of the defendant confessing, despite being innocent: $P(r \mid \overline{g})$.

Similarly, the *Defendant Fallacy* occurs when the evidence m is considered to be unimportant, when a very high prior for $P(\overline{g})$ (due to a large number of potential suspects), still yields a large $P(\overline{s} \mid m)$.

Moving a bit up the chain, we find the $P\,(\textit{Another Match})$ *Error*, which is committed when one equates the the value of $P\,(m \mid \overline{s})$ to the probability that *at least* one innocent member of the *reference population*[1] has matching evidence. Suppose that a certain trait that characterises a sample of DNA is found in one out of X members of the reference population, which has size N. Then, the probability of a randomly selected person from that population matching the evidence found on the crime scene is $1/X$. However, the probability that there is *at least* one person, other than the defendant, in the reference population that will match the evidence found at the crime scene, is $1 - (1 - 1/X)^N$.[2] Clearly, the size of the reference population is relevant for this estimate, while, according to Koehler, there is little awareness amongst experts, attorneys and judges about this fact.

A similar fallacy is the *Numerical Conversion Error*, where the value of $P(m \mid \overline{s})$ is confused with the number of other people from the reference population that would have to get tested to find someone who matches the evidence.[3]

Another related fallacy is the *Expected Values Implying Uniqueness* fallacy, where it is assumed that if the size of the reference population is of order $1/P(\overline{s} \mid m)$, the defendant must be the *only* match. However, it can be shown [3] that the chance that there will be at least two matches in a population of size $1/P(\overline{s} \mid m)$ exceeds 25%.[4]

Another fallacy related to the reference population is the *Defendant's Database Fallacy*. Here, the value of $P\,(\overline{s} \mid m)$ is based on a different population from that determined by $P(s)$ or $P(g)$. This can occur if, for example, the defendant is found because a DNA sample found at the crime scene matches the DNA of the defendant found in a database of persons with earlier convictions, and not because there was other evidence that led to suspicion towards the defendant. The the value of $P\,(\overline{s} \mid m)$ might then be based on the probability of a random member from the database matching the sample, while $P(g)$ may be based on all persons that were within five kilometers of the crime scene during the day at which the crime took place.

A very obvious, though common fallacy is that of *Base Rate Neglect*, or failing to take into account prior odds such as $P(g)$ or $P(s)$. In general, this might lead to overestimations of an event occurring when the event is more unusual than it seems, or underestimations if the event is less unusual than it seems.

One particular class of probabilistic fallacies is known as the *Transposed Conditional Fallacies*. Here, a probability of an event that is conditioned on certain evidence is confused with the probability of that evidence, conditioned on the event.

Equating $P\,(m \mid \overline{s})$ with $P\,(\overline{s} \mid m)$ is known as the *Source Probability Error*, because the probability of the evidence matching the defendant if the defendant is not the source, is confused with the probability of the defendant not being the source, if a match is found for the defendant and the evidence.[5] This tends to exaggerate the strength of the evidence. In particular, in order to compute $P\,(m \mid \overline{s})$, we need the prior for $\overline{s}$, and we thus need information about the size of the source population. Alternatively, this fallacy is often known as the *Prosecutor's Fallacy*.

Another fallacy that sometimes goes by the name of *Prosecutor's Fallacy* is the *Ultimate Issue Error*, when it is implicitly assumed that $P(g) = P(s)$, on top of committing a source probability error. Thus, someone committing an ultimate issue error will incorrectly assume that $P(m \mid \overline{s}) = P(\overline{g} \mid m)$. Another type of ultimate issue error is committed when $P(m)$ is assumed to be equal to $P(g)$, and it is thus concluded that $P(m \mid \overline{s}) = P(\overline{g} \mid r)$. This would mean that the probability of the defendant matching the evidence if the defendant is not the source of the evidence, is equal to the probability of the defendant not being guilty when the lab reports that a match is found.

3.2 Fallacies in combining evidence

The fallacies described above all stem from a misunderstanding of conditional probabilities in dealing with single pieces of evidence. Another type of fallacy is committed when a person fails to combine multiple pieces of evidence together correctly.

The first, and very obvious, fallacy is that of the *Dependent Evidence*, also known as *Double Counting*. This fallacy is committed if someone treats two pieces of evidence that are dependent of one another as if they were independent. A famous example is that of Sally Clark. Two of her babies died of sudden infant death syndrome (SIDS), and those deaths were treated as independent.

[1] The reference population is used to determine the *random match* probability $P(m \mid \overline{s})$, the probability that a person not involved in the crime, coincidentally provides a sample that matches the evidence. Depending on the information available about the source of the sample found at the crime scene, this reference population can consist of the general population, or can be a case-specific *source population*. For example, if an eye witness has provided information about seeing a caucasian male, this could exclude many members of the general population from the source population. Note that there is a subtle difference between the *source population* and the *potential suspect population*. For example, when woman is murdered in her bed, only a week after her husband died, her husband is a member of the source population, but not of the potential suspect population.

[2] Koehler provides an example in [10], with $X = 705,000,000$ and $N = 1,000,000$. In that case, $P(m \mid \overline{s}) = 1/705,000,000$, but the probability that at least one person other than the defendant can provide a sample that matches the evidence at the crime scene is roughly $1/714$.

[3] In the previous example, we would not have to test 705,000,000 people to have a decent chance at finding a match, but rather a number n such that $(1 - 1/705,000,000)^n \leq 0.5$, which corresponds to testing roughly 489,000,000 people.

[4] We can use the Poisson distribution for computing the probability of seeing x matches, if the expected number of matches is λ. This probability is given by $P(x \mid \lambda) = \lambda^x e^{-\lambda}/x!$. In our case, since the probability of a random innocent person from the reference population having a match is $P(\overline{s} \mid m)$, and the population has size $1/P(\overline{s} \mid m)$, we conclude that $\lambda = 1$. This yields $P(x > 1 \mid 1) = 1 - (P(0 \mid 1) + P(1 \mid 1)) = 1 - 2e^{-1} \approx 0.26$. Note that we have not used any information about the size of the reference population in particular.

[5] An example: suppose a scientist finds that if an animal is a cow, there is 100% chance that it has four legs. When you commit a source probability error, you take this to mean that if an animal has four legs, there is a 100% chance that it is a cow.

According to a medical expert witness, about two out of 17200 babies in families like Clark's die of SIDS. Squaring that number yielded a probability of roughly one in 74 million that two children would die of SIDS in one family. In 2002, the President of the Royal Statistical Society in the UK wrote a letter to the Lord Chancellor to point out the fact that, in order to assume that cases of SIDS occur independently in families, one has to prove this in court, which had not happened. Also, there are reasons for assuming that cases of SIDS do not occur independently in families, there might be genetic or environmental influences that predispose some families to SIDS.[6] A similar mistake was made in the case of Lucia de Berk. In calculating the probability of there being six suspicious incidents during her shift in one year, it was assumed that all nurses and all shifts were interchangable with respect to the possibility of a person dying during a shift. Although medical specialists tend to claim that this is in fact the case, nurses tend to disagree. In their experience, patients tend to die on the shifts of nurses they feel comfortable with. On top of that, events marked as 'incidents' in the Lucia de B case, always started with a phone call from a nurse to a doctor. Different nurses have different personalities, and their decision to call a doctor to report an incident is also based on personality traits like self-confidence, as well as professonial and personal experience and attitude [12].

This leads to a special case of the dependent evidence fallacy: the *Logically Dependent Evidence Fallacy*, where one piece of evidence follows logically from another. If the two pieces of evidence are that a particular nurse makes a lot of phone calls to report incidents, and that that particular nurse has a lot of incidents on her shifts, these are logically related.

The *Conjunction Fallacy* is a typical product of human limitations. It stems from the difficulty that humans have to retain different pieces of information in memory, which they solve by merging new pieces of information with the old ones. In this process, certain subtleties such as the uncertainty in a piece of information, tend to get lost. Therefore, people might fail to take into account the different uncertainties that a piece of evidence is made up of, and thus assign a larger weight to the piece of evidence than it should have [14].

The *Coincidence Fallacy* occurs when an observed combination of events is implied to have a very small probability, as was done in the Lucia de Berk case. In this particular example, people failed to see the great variance of the number of suspicious incidents on shifts of a particular nurse within one year, thus judging that a large number of incidents is very, very unlikely (the famous one in 342 million number). They were proved wrong in [7, 12]. This fallacy leads to an underestimation of the probability of such an observed combination of events. For example in [12], it is argued that, if we were to consider the yearly number of suspicious incidents during Lucia's shifts as strong evidence for her being a serial killer, we should send one in twenty-six nurses to jail each year, or maybe even one in nine! This means that the yearly number of suspicious incidents during Lucia's shifts is not that unusual.

A tricky fallacy is that of *Previous Convictions*. It is a special case of the *Jury Observation Fallacy*, and can be illustrated as follows. Suppose that a jury, in a serious crime case, has found the defendant to be not guilty. This is told to another person, who has nothing to do with the case. Next, this person is told that the defendant was earlier convicted for a similar crime. Now, the person is asked if this information increases or decreases his/her belief in the correctness of the jury's verdict. Typically, this person's confidence in the correctness of the jury's verdict will decrease after being confronted with the extra information of the defendant being convicted earlier for a similar crime. The tricky part is that it can be shown that the probability of the defendant being guilty in the second case actually *decreases* once the prior conviction is known [4].

Finally, there is a fallacy that is called *What You See Is All There Is* (WYSIATI), by Daniel Kahneman in [8]. On the one hand, this expressed as investigators jumping to conclusions on the basis of limited evidence, but in some sense, it can also manifest itself as investigators failing to see the *absence* of evidence [14]. To illustrate this point, consider the story called *Silver Blaze*, in Sir Arthur Conan Doyle's *The Memoirs of Sherlock Holmes*. It contains the following dialogue between Inspector Gregory and Sherlock Holmes:

"Is there any point to which you would wish to draw my attention?"

"To the curious incident of the dog in the night-time."

"The dog did nothing in the night-time."

"That was the curious incident," remarked Sherlock Holmes.

From this, Holmes infers that a member of the household must have been the culprit, as the dog would have barked to a stranger coming into the house.

4 Biases in estimating subjective probabilities

Above, we have discussed several common fallacies that people commit when confronted with reasoning with uncertainty. As we have seen in Section 2.2, we do not only have to reason about (in)dependent events in a proper way, we also have to assess certain probabilities, that might be subjective. Suppose, for example, that a forensic expert on facial recognition is given a photograph of the defendant, and a picture taken of the perpetrator of a crime. Suppose that the defendant has a very clear mole and scar on his/her right cheek, and that these features are also visible in the photograph of the perpetrator. The expert will compute the LR of seeing the mole and scar in the picture of the perpetrator, given that the defendant is or is not the perpetrator. To obtain posterior odds, the prior odds have to be assessed first. Therefore, an assessment must be made of size of the suspect population, and the number persons in that population that have a mole and a scar on their right cheek. Chances are that there are no statistics about this last trait, so these numbers have to be estimated subjectively [11].

[6] Incidentally, more fallacies were committed in the Sally Clark case. The prosecutor's fallacy led people to conclude that if the probability of two children dying of SIDS in one family is one 74 million, then that is the chance that the deaths were indeed accidental. This is a fallacy because by only looking at the probability of the two deaths being accidental, one does not take into account the prior probability that both children were murdered, which might be equally small, if not smaller. Recall from Section 2.2 that considering an alternative hypothesis is the bread and butter of reasoning with uncertainty in court.

Note that we define *objective probabilities* to be values that are calculated according to the laws of the probability calculus, for example, Bayesian reasoning. *Subjective probabilities* are those that are estimations of the probability of an event, given by a person ('subject').

In this section, we highlight a number of common biases and heuristics with respect to estimating subjective probabilities. These biases allow humans to deal with probability assessment in a quick way, with limited mental effort. At times, this is useful, but it can also be inaccurate, and tricky to avoid.

4.1 Representativeness

The heuristic we want to highlight is that of *representativenes*, this section is based on [9, 17]. When assessing probabilties, people judge the probability of an event occurring by judging the extent to which the event is representative for what they expect. For example, suppose someone tells you about their former neighbour, Steve [17]. Steve is described as "very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail." Now, you are asked to rank a list of occupations (for example, farmer, salesman, airline pilot, librarian or physician), according to your estimation of the probability that Steve is engaged in such an occupation. Before you read on, we suggest you grab a pen and draw the list for yourself.

Typically, people will rank their occupations according to how well the mental image created by the description fits the mental stereotype they have of a farmer, a salesman, etcetera. In other words: the degree to which the description of Steve is representative for a farmer, a salesman, etcetera, determines how likely we think him to be engaged in those occupations. People consistently judge the more representative event to be more likely. This is a form of *substitution*, when faced with a hard question ("Which occupation is more likely?"), our brains substitute it with a different question ("Which stereotype is most similar to the description?"), and we answer that question instead, thus reducing the mental effort needed for completing the task [8].

So what is representativeness? In [9] the representativeness of an event is defined as the degree to which it is *similar* in essential properties to its *parent population*, and the degree to which it *reflects* the salient features of *the process by which it is generated*. We give two examples to illustrate these criteria.

Consider the following problem [9]: "All families of six children in a city were surveyed. In 72 families the *exact order* of births of boys and girls was GBGBBG. What is your estimate of the number of families surveyed in which the *exact order* of births was BGBBBB?" Assuming the birth of a boy or a girl to be equally likely, both birth sequences are equally likely. However, the first one is more representative than the second one, since it reflects the proportion of boys and girls in a population. Thus, people judge the first sequence to be more likely than the second. This is an example of the first criterium; boys and girls in equal proportions is what you expect in the total population. The second criterium can be illustrated by asking the same question, only for the sequences BBBGGG and GBBGBG.

Here, the first sequence seems less likely, as it doesn't reflect the randomness we expect to see in the births of boys and girls, and thus does not reflect the features of the process by which birth sequences are generated. As a consequence, only a subset of all equally likely possible outcomes is perceived as a representative sample.

Having defined the scope of the representativeness heuristic, we discuss a few errors and biases that are associated with it.

Firstly, the representativeness heuristic causes an *Insensitivity to Prior Probability of Outcomes*, as different priors have no effect on representativeness, but they do have an effect on the (posterior) probability of an event. To illustrate this, recall Steve. For most of us, the description of Steve matches the most with our stereotype of a librarian. However, as there are more salespersons than librarians, this fact should enter into the equation to get a reasonable estimate of the probability that Steve is a librarian, versus that of Steve being a salesman.

This effect is clearly present, even when subject are aware of the prior probabilities, and know what they mean and how to reason with them. Let us consider an example that is similar to the one described above [17]. Suppose we have a population of 100 men, 70 of them are engineers, while the other 30 are lawyers. Now, subjects are confronted with the description of a person drawn from that population, and they have to estimate the odds that that description belongs to an engineer or a lawyer. When provided with no description, subjects correctly estimate the probability of a person being an engineer to be 70%, and that of him being a lawyer to be 30%. However, this changes when provided with a description like "Dick is a 30 year old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues." This description contains no information that makes it to be more representative of a lawyer or of an engineer, so essentially, we are given no information. However, subjects tend to interpret this as meaning that there is a fifty-fifty chance of Dick being an engineer or Dick being a lawyer, and forget about the prior probabilities for these occupations in the population.

Aparrently, people respond differently when they are given no evidence, than when they are given worthless evidence. This is particularly interesting in the light of the subject of this paper: reasoning with probability in court.

Another mistake made due to the representativeness heuristic is that of *Insensitivity to Sample Size*. This is illustrated in [9], where the authors asked subjects to come up with probability densities for three types of distributions: a binomial distribution with $p = .5$ for each of the two outcomes, a binomial distribution with $p = .8$ for one of the two outcomes, and a distribution of height. They let subjects draw probability distributions for populations of sizes 10, 100 and 1000, and found that the median distribution given by the subjects was independent of population size. In reality, one would expect the distributions to be 'flatter' for small populations and more localised around p (or the average height) for larger

[7] This can be further illustrated with an example [17]. Suppose we have an urn filled with balls, 2/3 of which are of one color, and 1/3 of which are of another colour. Suppose John draws five balls from the urn and finds that four are red and one is white. Jane draws twenty balls, twelve of them red and eight of them white. Who should feel more confident that 2/3 of the balls in the urn are red, and 1/3 of the balls are

populations.[7]

A mistake that is related to the one of insensitivity to sample size is that of the *Misconception of Chance*, or the belief in the *Law of Small Numbers* [16]. It is not limited to laymen, also. This mistake is made when people assume a small sample to have a high probability of showing a trend that is representative for a large sample. For example, people expect to see high degrees of randomness in small sequences of random events, such as the tossing of a coin. This is also known as the *Gambler's Fallacy*, when roulette players expect black to be due after a sequence of reds, and can thus manifest itself as an expectation that people have that chances will 'cancel eachother out' in short sequences, because that is what they do in the long run. This expectation of *local representativeness* leads us to underestimate the probability of extraordinary samples when the samples are small.

Another interesting mistake that is being made is that of *Insensitivity to Predictability* [17]. Let us get back to Steve. The description of Steve was given by a former neighbour of his, and we have no information on how well this neighbour knew Steve, how long ago it was that they were neighbours, how skilled the neighbour is in judging personalities, etcetera. Did you take this into account when ranking your list of possible occupations of Steve according to the probability that he is engaged in such an occupation? You probably did not. Even people trained in statistics will perform this task based on representativeness, rather than actually assessing the likelihoods based on the predictive value of the description [8].

4.2 Availability

An interesting heuristic is that of *availability*, in which we assess the probability of an event happening based on the ease with which an example of such an event springs to mind. This ease can be due to different effects.

The bias of *Retrievability of Instances* is the most direct way in which examples of an event are available in the forefront of our brains. Events are easier retrieved if they occured recently, explaining why we are more anxious of being the victim of a terrorist attack after there just has been one. Events are also easier to retrieve if they are in some way familiar, for example, when they happen to family members or friends, or even famous people. Having once seen a house burn down also makes it easier for us to retrieve an example of a house burning down than only having read about houses burning down; personal experience also influences the ease with which we retrieve examples.

A maybe less obvious bias is that of the *Effectiveness of the Search Set*. When confronted with the question as to whether there are more words in the English language that start with an r (such as 'rain' or 'river'), or more words that have an r at the third position of the word (such as 'more' or 'word'), we find it easier to recall words that start with an r, than to recall words that have an r at the third positions. We thus tend to conclude that there are more words in the English language that start with an r, than that there are words with an r at the third position, while in fact words with an r at the third position are more frequent than those with an r at the first position.

These two biases again illustrate that we tend to judge events that take less mental effort to conceive to be more probable than events that require our brains to work harder to imagine them.

4.3 Adjustment and Anchoring

Anchoring is providing a person with a number that can be used as a starting point for assessing a subjective probability. The person can *adjust* this number until he or she thinks it to be appropriate. However, as it turns out, people generally do not adjust the number enough, since they intuitively feel that the true number cannot be too far away from the initial one they have been given, which is yet another heuristic humans use when assessing probabilities.

This knowledge is very powerful if you are often required to haggle over a price. Starting out with proposing a very low (or high, depending on whether you are buying or selling) can be very lucrative. In the context of buying or selling, however, you do have to be reasonable. The price you propose is always based on something, and you have to be able to defend it to some extend. The surprisingly, the *Insufficient Adjustment* bias also works when the anchor you have been given is obtained randomly, and should thus contain no information about the true value. Tversky and Kahneman report an experiment in which subjects were asked to estimate the number of African countries in the United Nations [17]. The subjects were divided in groups; and each group got a different anchor, that was obtained by spinning a wheel of fortune. The median of the guesses for the number of African countries in the UN from the group that received 10 as an anchor, was 25. The group that got 65 as an anchor has a median of 45 in their guesses.

A different manifestation of the use of the anchoring and adjusting heuristic is seen in subjects estimating *conjunctive and disjunctive events*. People tend to have a bias towards overestimating the probability of conjunctive events, and underestimating the probability of disjunctive events. The overall probability of a conjunctive event is lower than the probability of a singular event, while the overall probability of a disjunctive event is higher than that of a singular event. When anchored with the probability of a singular event, which is always a natural place to start when considering conjunctive or disjunctive events, people do not adjust the conjunctive or disjunctive probabilities sufficiently to get to a correct estimate. In the context of a court case, in which several pieces of evidence may have to be combined, this may have significant effects.

However, this might be compensated by us knowing in which direction the bias will occur. In the case of guessing the numbers of African countries in the UN,

white? Intuitively, most people will reckon that John has much stronger evidence of having drawn from an urn with $^2/_3$ red balls and $^1/_3$ white balls than Jane. However, the correct posterior odds are: $\frac{P(2/3 \text{ red}|4r,1w)}{P(1/3 \text{ red}|4r,1w)} = \frac{P(4r,1w|2/3 \text{ red})}{P(4r,1w|1/3 \text{ red})} \times \frac{P(2/3 \text{ red})}{P(1/3 \text{ red})} = \frac{(2/3)^4 \cdot 1/3}{(1/3)^4 \cdot 2/3} \times \frac{.5}{.5} = \left(\frac{2/3}{1/3}\right)^3 = {}^8/_1$ for John, and similarly $^{16}/_1$ for Jane. Therefore, Jane is more certain than John of having drawn from an urn containing $^2/_3$ red balls. John's set of drawn balls provides stronger intuitive evidence of being drawn from an urn with a majority of red balls than Jane's set of balls, since the proportion of red balls to white balls is larger in John's draw (4:1) than in Jane's (12:8, or 3:2). This is an illustration of how the judgment of posterior odds is dominated by the extent to which a sample has proportions that are consistent with what we expect, and almost unaffected by the size of the sample.

we can never be sure if people will adjust upwards from the anchor, or downwards. With the bias in estimating conjunctive and disjunctive events, the structure of our problem will give us a hint about the direction of our bias, which helps us in reducing that bias.

There are many more biases due to the three heuristics described above, but discussing these is beyond the scope of this work.

5 How to interpret and design Bayesian models?

Although Bayesian reasoning is an established method for reasoning with uncertainty, which is well documented and understood, there exist different ways of solving the same problem, and Bayesian Networks that represent a certain Bayesian argument need not be unique to yield acceptable results. In this section, we discuss how different ways of thinking about Bayesian reasoning affect the choices we make and how we interpret the result.

5.1 Three Bayesian semantics

In [15], Shafer and Tversky distinguish three distinct semantics for the Bayesian 'language' that describes how to deal with reasoning with uncertainty according to Bayesian principles.

The first semantics is that of *frequency*. We look at our evidence by asking how often, in a repeated experiment consisting of the situation at hand, the truth will turn out the way it does for our evidence. The second semantics is that of *propensity*. Here, we view the evidence in the context of a causal model, and ask about the propensity of the model to produce the results we observe. Note that, while the last view seems a very natural way of thinking for us, it might predispose us to commit certain fallacies and fall for certain biases more often than when looking at our model through frequency eyes. For example, the propensity approach could make us more vulnerable for the base rate neglect fallacy, or the insensitivity to predictability bias. On the other hand, the propensity view might cause us to spot dependencies more easily, thus being less vulnerable for fallacies in combining different pieces of evidence.

Finally, the third semantics is that of *betting*. Viewing our evidence in the light of betting, we might make comparisons of the probability of certain events by assessing our willingness to bet a certain amount of money on it. This view might be the least useful in interpreting Bayesian reasoning, as several studies have shown that people are often willing to loose money, in order to continue to believe in what they believe [15]. Also, people tend not to derive believes from their bets, but bet on what they believe in. On the other hand, the betting semantics is a more general view of probability than the frequency and propensity semantics. However, the betting semantics does not help us much in our understanding of the probabilistic principles in our model.

5.2 Two Bayesian designs

Besides recognising three semantics with which we can look at our evidence and model, Shafer and Tversky distinguish two main ways of organising our model: the *total evidence design* and the *conditioning design*.

The total evidence design bases each probability and each conditional probability on all the evidence that is available. This is used for computing the probability of certain outcomes. Note that this ordering of events is opposite from how we tend to think about evidence in a court case. Recall from the chain of reasoning in Figure 1 that we start with the event (defendant is guilty or not), and work towards the probability of this leading to us finding the evidence that we found. This can be turned around by considering an alternative chain of reasoning (see Figure 2), that does not have the causal nature of the chain as presented in Figure 1, but expresses a similar idea [10].

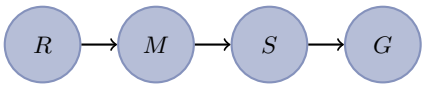

Figure 2: An alternative way of seeing the causal chain of events of Figure 1, now expressed as a chain of reasoning, rather than a chain of events. We might get a reported match between the evidence and a sample from the defendant (M). This is suggesting, although not guaranteeing, a true match between the evidence and the defendant (M). This is again suggestive (but by no means certain) of the defendant being the source of the evidence (S), which in turn is suggestive of the defendant being guilty of committing the crime (G).

Please note that for more complicated examples than that of Figure 2, there might be different choices for organising our evidence. A limitation of this design is that its effectiveness depends on the decomposability of the evidence. This in turn depends on how the information we have is organised, which both reflects and determines how we think about the problem.

For cases in which the evidence is less clearly decomposable, an alternative approach is to choose a conditioning design of our Bayesian model. Here, we build a model with a probability distribution, based on a hypotheses. Then, we condition the model on certain pieces of evidence, to see what the probability is of us finding our other pieces of evidence. If the probability of finding these other pieces of evidence is low, this indicates that our model is probably based on a hypotheses that is unlikely to be true. While the name of this design attracts our attention to the second part of the process, it is the model building and assessing of the probability distributions that is the hard part.

Within the condition design of a Bayesian model, there are again two directions to choose from; an *observational design* and a *partitioning design*.

In the observational design, the evidence whose probability of finding is determined by conditioning our model on the other evidence, is deliberately obtained *after* the model is built. This has as advantage that it might yield more open minds while building the model, but it might also be wise to let the new evidence be obtained by a person who is unfamiliar with the model. This approach might have been benificial in the Lucia de Berk case. It could have been used to assess how likely it is to find a certain number of suspicious incidents happening on the shift of the same nurse within a year. However, this would take some time (at least a year), so even if it is conducted properly, it might not be very helpful. Another approach would be to just not use the records of suspicious incidents while building the model, and consulting them only after building the model.

In the partitioning design, the total evidence is partitioned in 'old evidence' and 'new evidence,' and the probabilities are assessed based on the old evidence, after which the model is conditioned on the new evidence.

This section has given a quick overview of some of the design choices that can be made while constructing a Bayesian model for reasoning with uncertainty. For a more detailed description, please refer to [15].

6 The role of Bayesian reasoning in court

As we have seen above, reasoning with probabilities comes with all sorts of complications. We must watch out for the probabilistic fallacies that prey on us when we reason with uncertainty. Furthermore, we must have the discipline to really think about the (implicit) assumptions we make when estimating subjective probabilities. On top of that, we have to have a good understanding of what we are doing when organising our evidence in a Bayesian Network.

How can we use all this in court? What ways are there to organise the responsibility for these different aspects of reasoning with uncertainty? What are the problems with these different approaches? In this section, we highlight ways of dealing with reasoning with uncertainty in court, using Bayesian theory in particular.

6.1 Bayesian reasoning in court

In [13] it is argued that, as lawyers are taught a certain type of statistics, they have a limited view of the concept of probability. To illustrate this, they quote a judge:

"The concept of 'probability' in the legal sense is certainly different from the mathematical concept, indeed it is rare to find a situation in which these two usages coexist, although when they do, the mathematical probability has to be taken into the assessment of probability in the legal sense and given its appropriate weight."

A main limitation of the statistical teaching to those practicing law, is that it argues that axioms of probability apply only to repeated experiments, thus suggesting that there exists no rational way of assessing certain case-specific probabilities, hence the quote above.

Yet, the authors of [13] argue that expert witnesses should *only* testify to the LR of the evidence. The assessment of the prior probalities of the hypotheses, and thus that of the posterior probabilities, should be left to the court. This is also how evidence is used in the Dutch system [11]; the expert witness provides a testimony about the LR of the two hypotheses that he/she has investigated with respect to the given evidence. Then, the judge has to assess the prior odds of each piece of evidence, and combine the posterior odds of all pieces of evidence to come to a judgement. In the Dutch system, the NFI (Nederlands Forensisch Instituut, or Dutch Forensic Institute) provides a report for each piece of evidence, in which the LR is not only given in numbers, but also verbalised, in a way that is shown in Table 1.

Table 1: The verbal terms of probability and their numerical definition, as used by the NFI [11].

Order of magnitude of $LR = \frac{P(\mathbf{results}\|H_1)}{P(\mathbf{results}\|H_2)}$	**The results of the test are...**
1 – 2	about as likely
2 – 10	a bit more likely
10 – 100	more likely
100 – 10,000	much more likely
10,000 – 1,000,000	very much more likely
> 1,000,000	extremely more likely
	... when H_1 is true than when H_2 is true.

They note that an advantage of this approach is that probabilities can still easily be communicated when there are no exact numbers. If an expert can only give a rough estimate, they can use the verbalised terms of probability to still communicate likelihoods in a standardised fashion. Furthermore, each report on a piece of evidence is accompanied with an attachment that explains the basics of Bayesian reasoning, two main fallacies (the prosecutor's fallacy and the source probability error) and contains Table 1.

In such a system, we can expect judges and lawyers to be familiar with Bayesian reasoning, and we can expect them to have experience in using Bayesian reasoning properly and avoiding fallacies. As we have seen with the Lucia de Berk case, this expectation might be unjustified or naive. In [5], Norman Fenton and Martin Neil also argue that all fallacies mentioned in Section 3 can be easily avoided by having just some understanding of Bayes' Theorem. However, in a system such as the one in the United States, where a jury has to make a judgement, we cannot expect members of a jury to fully understand Bayes' Theorem. This sentiment is shared by the people actually in court, as can be shown by the following quote from an appeal judge [5]:

"The introduction of Bayes' theorem into a criminal trial plunges the jury into inappropriate and unnecessary realms of theory and complexity deflecting them from their proper task. The task of the jury is ... to evaluate evidence and reach a conclusion not by means of a formula, mathematical or otherwise, but by the joint application of their individual common sense and knowledge of the world to the evidence before them."

While the first part of this statement is something we might feel sympathetic to, the second part is worrying, since we know from Sections 3 and 4 that people in general and laymen in particular are very vulnerable to making mistakes in working with probabilistic evidence, and thus *cannot* be trusted to rely on common sense and still come to the right conclusions. Therefore, Fenton and Neil argue that Bayesian reasoning must be used in court by the jury, and they propose to use Bayesian Networks to help the jury make their judgement.

6.2 The use of Bayesian Networks in court

Fenton and Neil acknowledge the problem of communicating abstract models of probability to a jury, so they propose to ease the members of the jury into drawing their own (correct!) conclusions about statistical problems in small examples. Then, they feel the jury is sufficiently equipped to deal with more complex

Bayesian reasoning. To support this claim, they argue that Bayesian reasoning can be seen as something similar to long division:

1. We understand the basic principles behind it, and we can work out small examples by ourselves;
2. Scientists have developed algorithms for the general case;
3. The algorithms do not need to be understood by laypeople, as sufficient experts have tested and validated them;
4. There are tools that implement the algorithms to acceptable degrees of accuracy (a calculator in the case of long division, a tool like SamIam [1] in the case of Bayesian reasoning);
5. There are different tools implementing similar algorithms and they all give approximately the same result;
6. Most people are able to enter the basic assumptions into the tool and press the relevant button to get a correct result (for long division, the basic assumptions consist of determining what number you want to divide by what other number, for Bayesian reasoning, the assumptions are a bit more complex).

In this section, we first show how the jury might be taught about Bayesian reasoning in small examples, according to Fenton and Neil. Then, we discuss in short the validity of the six assumptions made above for Bayesian reasoning.

Fenton and Neil argue that a visual representation of the probabilistic model is benificial in helping laymen to make correct probabilistic computations themselves. They propose to either visualise them by a simple animation of 'stick figures,' where parts of the population are highlighted, according to the probabilities they represent. An alternative approach is to use an *event tree* of the kind presented in Figure 3.

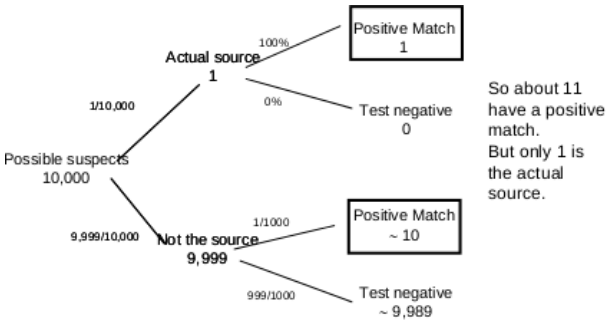

Figure 3: An example of an event tree that will help jurors gain insight into Bayesian reasoning. Figure from [5].

In that particular event tree, there is a population of possible suspects of 100,000. Assuming that we are looking for one perpetrator, the prior probability of the defendant being the actual source of the evidence found at the crime scene is thus $^1/_{100,000}$, while the prior probability of the defendant not being the source is $^{99,999}/_{100,000}$. If, for the sake of simplicity, we make the assumption that $P(r \mid s) = P(m \mid s) = 1$ in terms of the causal chain of Figure 1, we are certain to get a reported positive match if the defendant is the source of the evidence. However, there is also a probability of getting a postive match if the defendant is not the source, and thus a probability of the defendant being innocent, despite being a positive match. This example helps to explain to jurors the prosecutor's fallacy, and thus helps to teach them to avoid it.

Now, Fenton and Neil assume that, having explained this example to the jurors and having them come to their own conclusions regarding possibility, they have checked off the first assumption given above. Let us assume that the jurors agree to accept assumptions 2 — 5. Now comes the tricky part. Fenton and Neil propose that an expert witness constructs a Bayesian model that summarises the evidence in the case. This expert also constructs Bayesian network based on this model, together with the prior probabilities for each variable in the model. In court, the jurors will watch while the expert enters the found evidence into the model, watching the probability of the defendant being guilty change as variables are conditioned on with the evidence. Figure 4 shows a simple example of what this might look like.

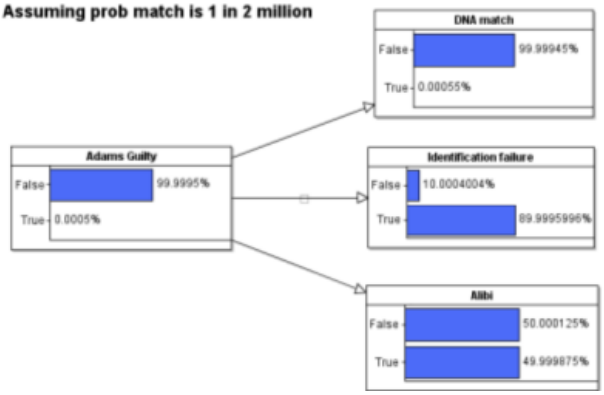

Figure 4: An example of a Bayesian Network as used in a court case. Three pieces of evidence are taken into consideration: a match for the DNA of the defendant (Adams) with a DNA sample from the crime scene, a witness failing to identify the defendant and the existence of a strong alibi. The network starts out with only prior proabilities, and the evidence is added one by one by conditioning on those variables. While doing so, the probabilities of Adams being guilty change. Figure from [5].

Fenton and Neil claim that all the jury has to do, is judge whether or not the assumptions made by the expert are reasonable, and, given the result, whether or not the defendant is found guilty. This will only work if assumption 6 is true: if "most people are able to enter basic assumptions into the tool and press the relevant button to get a correct result."

This is where we consider them to be somewhat optimistic. Fenton and Neil's objective is to relieve the jury from having to deal with mathematical formula's, and to avoid them committing fallacies of the types described in Section 3. While the first objective is achieved by using a Bayesian Network to visualise the mathematics going on in the model, some fundamental issues are not solved.

For example, Fenton and Neil expect the jury to judge whether or not the choices made by the expert are reasonable or not. By definition, they should know about the fallacies described above, since otherwise they would not be able to determine whether or not the fallacies are committed. A solution is to ask another expert to verify this, but that has always been an option, when only being provided with the mathematical model. We conclude that not much is gained in this respect. On top of that, we have seen in Section 5 that there are many ways of constructing a Bayesian model and of interpreting it.

The jury has to have an understanding of these different designs and interpretations in order to decide whether or not the chosen structure is suitable for the problem at hand. Again, not much is gained by using a Bayesian network to visualise the model.

Next, the jury has to determine if the probability assessments of the expert are reasonable. This might be even more problematic when the jury is presented with a neat model in which they can tweak the parameters and get a visualised result immediately, than when they are confronted with a more abstract mathematical formula. The anchoring and adjusting heuristic described in Section 4.3 might play a crucial role here, risking the tendency of jurors to judge certain prior probabilities as reasonable or not, depending on how much they cause the probability of the defendant being guilty to change when the prior probabilities change.

Finally, although jurors are not confronted with abstract mathematical models, they will still have to invest quite some time and effort in understanding reasoning with uncertainty, which they might still experience as tedious and "not their task." The many nuances and subtleties of Bayesian reasoning are impossible to explain using the simple event tree of Figure 3.

7 Conclusion

In this work, we have investigated the role of Bayesian reasoning in court. We have explained fallacies, heuristics and biases that are involved in reasoning with uncertainty. We have addressed different interpretations and designs for Bayesian models, and how they might influence the way we look at the problem. Finally, we have discussed a proposal by Fenton and Neil for using Bayesian Networks rather than mathematical formula's to present Bayesian reasoning in court. We find that many of the difficulties with abstract formula's are not really relieved when using a visualisation of the model instead. We do see that visualisations may help a layman in understanding some principles behind reasoning with uncertainty, but question how much is gained by visualising the model for the jury. In particular, the risk of committing probabilistic fallacies is not necessarily reduced by using a visualisation of the Bayesian model that is used to combine the different pieces of evidence in a court case.

References

[1] Automated Reasoning Group of Professor Adnan Darwiche at UCLA. Sensitivity Analysis, Modeling, Inference And More. `http://reasoning.cs.ucla.edu/samiam/`, 2004–2010.

[2] Ton Derksen. *Lucia de B. Reconstructie van een gerechtelijke dwaling.* Veen Magazines, 2006.

[3] Ian W. Evett and Bruce S. Weir. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists.* Sinauer Associates Inc, 1998.

[4] Norman Fenton and Martin Neil. The "Jury Observation Fallacy" and the use of Bayesian Networks to present Probabilistic Legal Arguments. *Mathematics Today (Bulletin of the IMA)*, March 2000.

[5] Norman Fenton and Martin Neil. Avoiding probabilistic reasoning fallacies in legal practice using bayesian networks. June 2011.

[6] Richard Gill. Statistical errors in court: Murder by numbers. Lecture at TEDxFlanders, `https://youtu.be/cbkdhD6BsoY`, `www.math.leidenuniv.nl/~gill/RDG_TEDxAntwerpen_Naked.pdf`, retrieved on June 18, 2015, March 2014.

[7] Richard D. Gill and Piet Groeneboom. One in nine nurses will go to jail. *www.math.leidenuniv.nl/~gill/hetero2.pdf*, August 2007.

[8] Daniel Kahneman. *Thinking, Fast and Slow.* Doubleday Canada, 2011.

[9] Daniel Kahneman and Amos Tversky. Subjective Probability: A Judgment of Reprsentativeness. *Cognitive Psychology*, 3:430—454, 1972.

[10] Jonathan J. Koehler. Error and exaggeration in the presentation of DNA evidence at trial. *Jurimetrics*, 34(1):pp. 21–39, 1993.

[11] Nederlands Forensisch Instituut, Ministerie van Veiligheid en Justitie. Vakbijlage – De reeks waarschijnlijkheidstermen van het NFI en het Bayesiaanse model voor interpretatie van bewijs. `www.nederlandsforensischinstituut.nl/Images/nfi-vakbijlage-waarschijnlijkheidstermen-versie-2.1-oktober-2014_tcm119-571394.pdf`, October 2014.

[12] Piet Groeneboom Richard D. Gill and Peter de Jong. Elementary Statistics on Trial (the case of Lucia de Berk). *www.math.leidenuniv.nl/~gill/hetero6.pdf*, June 2010.

[13] Bernard Robertson and G. A. Vignaux. Don't teach statistics to lawyers! *International Conference On Teaching Statistics*, 5:542—548, 1998.

[14] D. Kim Rossmo. Criminal Investigative Failures: Avoiding the Pitfalls. *FBI Law Enforcement Bulletin*, September 2010.

[15] Glenn Shafer and Amos Tversky. Languages and Designs for Probability Judgement. *Cognitive Science*, 9:309—339, 1985.

[16] Amos Tversky and Daniel Kahneman. Belief in the law of small numbers. *Psychological Bulletin*, 76(2):105—110, 1971.

[17] Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185:1124—1131, September 1974.