STATISTICS AND PROBABILITY IN CRIMINAL TRIALS

THE GOOD, THE BAD AND THE UGLY

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PHILOSOPHY
AND THE COMMITTEE ON GRADUATE STUDIES OF
STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Marcello Di Bello

December 2013

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Krista Lawlor, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Thomas Ryckman

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Brian Skyrms

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Robert Weisberg

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

George Smith

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

ABSTRACT

Is a high probability of guilt, in and of itself, enough to convict? There are *prima facie* arguments for both a positive and a negative answer. Since the certainty of guilt is unattainable, one can argue that a high probability should be enough to convict, for if it is not, conviction itself would be unattainable. On the other hand, a well-known hypothetical scenario suggests otherwise. Imagine a prison yard with one hundred prisoners and only one prison guard. One day, ninety-nine prisoners collectively murder the guard. After the crime, one prisoner is picked at random and tried. His probability of guilt is very high—i.e. it is 0.99 because the participation rate in the collective murder is 99:100—but it seems unacceptable to convict him on mere high probability.

I maintain that a high probability of guilt, in and of itself, is not enough to convict. I argue that the prosecutor's burden of proof does not only consist in establishing the high probability of the defendant's guilt; it also consists in (1) establishing guilt with a *resiliently* high probability, and in (2) offerring a *reasonably specific and detailed* narrative of the crime. In the prison yard scenario, these conditions are not met. Even though the prisoner's probability of guilt is high, it is not resiliently high, because new, and possibly exculpating, evidence could lower it dramatically. Second, in the prison yard scenario we are not offered any well-specified narrative of the crime—e.g. we are not told which role the prisoner on trial played in the killing of the guard. So the lack of resiliency and the lack of narrative specificity give grounds for our intuition that a conviction based on mere high probability is not acceptable.

My account has applications to current debates in both epistemology and legal scholarship. Many epistemologists hold that the high probability of a proposition, in and of itself, does not give us knowledge of that proposition. I agree. To make progress here, I suggest that we consider a notion related to, but different from, knowledge—i.e. the dual notion of *reaching a justified judgment* and of *justifiably withholding judgment*. I draw attention to two features of reaching/withholding judgment which bear a close relation to condition (1) above, i.e. the resiliency condition. The first feature is the *stability* of one's judgment against future, and possibly contradictory, evidence; the second

is the *burdensomeness* associated with withholding judgment. I argue that whenever our judgment about a proposition p won't be stable (given our current evidence), we should withhold judgment about p instead, so long as doing so is not unduly burdensome for us or for those affected by our withholding judgment.

As far as legal scholarship is concerned, my account can also help us clarify some pressing issues regarding the use of statistical evidence in criminal trials. We have seen a steady rise in the use of statistics in criminal cases, especially as a result of the discovery of DNA fingerprinting in the 1980s. Relatedly, the phenomenon of "big data" has made it easier to find statistics for trial purposes. Against this background, a question naturally arises. Are statistics alone enough to convict? For one, we do feel uneasy about them: statistics seem to lack "specificity" because they place the defendant in a group with others, and we do not want to be convicted because of what others did. On the other hand, many are less uneasy in convicting on DNA evidence. The reason for this is that DNA evidence is considered more specific to the defendant, even though, in the end, its probative value rests on statistical estimates. This suggests that we should be careful with wholesale dismissals of statistical evidence in criminal trials. My position is that, in some cases, a conviction *may* justifiably rest on statistics alone, and in other cases, it *may not*. A deciding criterion, among others, is given by condition (2) above—i.e. the specificity of the narrative being offered by the prosecutor. I show that statistical evidence can be particularly problematic whenever it fails to support a well-specified narrative of the crime. This conclusion, while still leaving room for statistical evidence in courts, refines the earlier worry that statistical evidence lacks specificity. The feature of specificity, however, should be understood as a feature of narratives, not of individual pieces of evidence. The motivation for this is that we cannot isolate pieces of evidence; we are better off considering the evidence more holistically.

ACKNOWLEDGMENTS

Five years ago, Krista Lawlor and I started a reading group on the applications of epistemology to evidence law theory. It was then that I started thinking about statistics, probability, and the law. Krista let me explore the literature freely, but when I got lost, she put me back on track. I could always stop by her office and think things through with her. Thank you, Krista, for being there since the beginning and until the end. You have been a wonderful advisor. Thank you for believing in this project even when I was discouraged, lost, and confused.

I ran into George Smith in the copy room of the philosophy department. The copy machine was broken and we started talking. We quickly discovered we had a common interest: evidence in court! He has contributed tremendously to this dissertation, and I owe him an enormous debt of intellectual gratitude. He is a careful and solid scholar, and a wonderfully sensitive human being. He has a relentless intellectual curiosity and a penetrating mind. He is a living statement that research and scholarship, at the highest level, are worthy enterprises. I thank you, George, for being such a generous source of intellectual and human support.

I thought my thesis should undergo the scrutiny of a legal scholar, so I turned to Robert Weisberg at Stanford Law. In the classes I took with him, I always enjoyed his imaginative way of thinking. Many think that the subject of criminal procedure is boring, but nothing can be further from the truth when you study it with prof. Weisberg. He has always been wonderfully gracious, supportive, and thoroughly insightful.

Tom Ryckman has rightfully given me a hard time with the toy examples contained in this dissertation. Philosophers in the analytic tradition have a fetichism for toy examples. Tom has kept my fetichism at check. Speaking with him has also been quite inspiring. He is a creative and synthetic thinker. Some of the ideas, insights, and examples in this dissertation come from him.

Brian Skyrms agreed to find some time for me. I am grateful he did. With his incisively concise comments, he has kept me on track. I suspect Brian is somewhat of a legal skeptic, and if the pages that follow can make him entertain the possibility that the

law is not so unreasonable and confused as it seems, I'll be happy.

Besides my committee, I have spoken with many people about my dissertation over the years. Rahul Chaudhri has seen me writing it in Palo Alto, San Francisco, Menlo Park, and even Bangalore! He has always been there to talk about my dissertation (whether he wanted to or not). He knows my dissertation better than I do, and he could have written it better than I did. David Hills, on various occasions, has graced me with his polymathic mind. I enjoyed speaking with him about DNA evidence, probability, testimony, resiliency, etc. It's hard to find something he does not know. James Garahan has read many chapters of this dissertation. Not only did he correct many stylistic flaws, but he also shared with me the fresh perspective of a lawyer and an intellectual who thinks disinterestedly about probability and the law. Thank you, Jim! Han van Wietmarschen and David Taylor, especially in the early stages, have engaged with my writing closely and have raised very thoughtful questions. Stefano Osnaghi has always criticized anything I had to say with vehement rigour and dear friendship. Mike Redmayne, Federico Picinali, Amit Pundik, and Mark Kelman have commented on some of the materials in this dissertation. Their comments were very helpful, though I could not address them all. Alberto Maria Benedetti, Mauro Grondona, Carlo Lottieri, and Daniel Ho influenced my thinking and writing. I should thank the participants in the PHIL 500 dissertation seminar for providing helpful feedback. Last, but not least, I want to thank the students in my Probability & Law class which I taught at Stanford in Winter 2013. They were all incredibly receptive and engaged with the material. Their questions, comments, and final papers helped me think through these ideas a great deal.

Sergio Galvan, Michiel van Lambalgen, Dick de Jongh, Martin Stokhof, Joost Breuker, Krister Segerberg, and Johan van Benthem did not think I was such a bad student and made it possible for me to pursue a PhD in philosophy. This has been the most wonderful gift. While at Stanford, Rahul has been my dearest friend with whom I shared miseries and glories, gossiping and literature, madness and sanity, friendship and anger, English and Hindi. Ji ha, mera dost! I enjoyed living in Rinconada with Pedro and David and the already too often mentioned Rahul. We were a good group. We cooked. We partied. We argued. And all such things that friends do. It is too bad that our ways parted, but I hope that one day we will be reunited in Rinconada under the shadows of heaven. Laura made me spin and told me many things I did not know. Alan, who apparently holds a UK passport, has been my sailing and skiing partner when we were not busy writing our dissertations. He has also been an excellent office mate until a tyrant bureaucrat at Stanford kicked us out. Christine taught me how one can have values in a world with no values. With Stefano and Marcia, I swam in the pristine alpine lakes of California. Han and Chiann have been good friends, and I will never forget when we climbed Half Dome at night or jumped from an airplane. Carlos and Sam can keep a house in order very well and occupy themselves with the Frege-Geach problem. Sara made me realize some of my ideas are not politically correct, which does not imply that they are incorrect. Katrina has been my reluctant Spanish teacher. Kathryn showed me how talent can take many forms. Katia kept my memory of Italy alive, though we really do not agree about politics. Lena drove me crazy. Alberto, who is one of the most efficient administrators I ever met, has made my life much easier at Stanford. Greg, Christy, and Eliza have seen me when it was the end, and all I was doing was ... writing my dissertation! They have been amiably gracious nonetheless. They even proofread parts of this dissertation.

Some accompanied me from other continents. Philip convinced me that what matters is here-and-elsewhere; that what is not is real; that *credo quia absurdum* is the most reasonable thing to say. My friends from high school—Paola, Alice, Federica, Emanuela, Giulia, Eleonora—all have jobs (and some even have kids!), though I have been anticipating the bankruptcy of Italy for quite some time. And you, Anjali, made me skip classes and travel the world. It has been all a matter of luck, all unpredictable, all highly improbable. It all falls out of the scope of this dissertation.

Alla mamma e al papà per esserci sempre

CONTENTS

INTRODUCTION

This dissertation explores the role of statistics and probability in criminal trials, in particular, how statistical evidence can be used and misused in court, and how the concept of probability can further our understanding of the legal fact-finding process. The unifying question can be put as follows: *Are we legally justified in convicting a defendant on the basis of mere statistical evidence if the probability of his guilt is sufficiently high?* To this question, people typically respond in different ways:

> 'Certainly yes!—It is all a matter of probabilities; no evidence can establish guilt with absolute certainty, whether it is statistical or not. If a sufficiently high probability of guilt is not enough for a conviction, will anything ever be enough?'
>
> 'Certainly no!—It is doubtful that guilt can be quantified probabilistically, and even if it could, a high probability of guilt, based on mere statistical evidence, cannot be enough to convict. Criminal trials are not about gambling with people's lives!'
>
> 'Wait!—I've never seen a defendant who has been brought to court on the basis of mere statistical evidence, so the question is a product of far-fetched speculation. Why should I indulge in mere speculation?'

Each response has its own merits and we can learn a great deal from understanding the underlying motivations. Chapters 1, 2, and 3 will survey the arguments and counter-arguments which can be found in the legal and philosophical literature of the last forty years.

The reader will be curious to hear about my own answer. I will show that in some cases defendants have been brought to court on the basis of mere statistical evidence;

chapter 4 contains some representative examples. So, contrary to what might appear at first, the question of this dissertation is not far-fetched speculation. Further, I will side with those who think that a high a probability of guilt, in and of itself, is not enough to legally justify a conviction. Although it leads to the conclusion that probability alone *is not* enough to convict, my argument will be probabilistic, and so it should appeal even to those who think that probability alone *is* enough to convict. If this sounds implausible, I invite the reader to have a look at chapter 2.

But if a high probability of guilt is not enough to justify a conviction, what could be enough? To answer this difficult question, I will offer an account of the criminal standard of proof 'guilt beyond a reasonable doubt.' A quotation from the Italian penologist Cesare Beccaria has served me as inspiration:

> It may seem odd that I talk of probability in relation to crime . . . [b]ut . . . moral certainty [*i.e. the criminal standard of proof*] is, strictly speaking, nothing but a probability, though a probability of such a sort to be called certainty. (Beccaria, 1764, cp. 14)

The moral certainty needed to convict is nothing but a probability, yet—as Beccaria rightly points out—this probability is also a certainty. This elusive remark suggests that the criminal standard of proof is a mixture of probability and certainty. What this mixture should be like, if anything at all, is a pivotal question in this dissertation. I will argue that in deciding whether to convict or to acquit, the probability of the defendant's guilt is one among several factors that should be taken into consideration. The other factors include the completeness of the incriminating evidence, the degree of specificity of the incriminating narrative, and the resiliency of the prosecutor's case against objections and challenges. These ideas are sketched in chapter 2 and then developed more fully in chapters 6, 7 and 8.

Here is a summary of what is to come:

Chapter 1—"*What Is Wrong with Statistical Evidence?*"—considers the worry that statistical evidence cannot by itself legally justify a conviction. A suggestion I explore in this chapter and throughout the dissertation is that statistics do not count as specific evidence because they place the defendant in a group with others. I argue that the notion of specific evidence, albeit promising, does not hold water. Instead of speaking of specific evidence, we are better off speaking of specific narratives.

Chapter 2—"*Is High Probability Enough to Convict?*"—looks at the debate between those who view probability theory as a tool to understand and conceptualize criminal trials and those who resist this idea. In particular, the chapter assesses the claim that an appropriately high probability of guilt, in and of itself, is enough to justify a conviction. I argue that this claim is mistaken because it views the criminal standard of proof as serving the function of error distribution while overlooking the equally important function of error reduction.

Chapter 3—"*A Debate that Began Forty Years Ago*"—reviews the literature on probability and statistics in criminal trial which appeared in the last forty years. This literature is mostly a response to Lawrence Tribe's seminal article *Trial by Mathematics* (1971). The scholarly debate, in fact, began two centuries earlier, but this thesis focuses on the contemporary legal and philosophical scholarship.

Chapter 4—"*How Statistics Get Used in Criminal Trials*"—reviews the pertinent case law. I examine three types of legal arguments which rely on statistical evidence for three different purposes: identifying the perpetrator of a crime; assessing whether or not an event occurred accidentally; and finally, estimating a total quantity when no direct measure of the total amount is available.

Chapter 5—"*Bayes in the Courtroom*"—offers an introduction to the mathematics and the philosophy of probability. It also gives a probabilistic, Bayesian analysis of the three types of statistical evidence that are discussed in chapter 4.

Chapter 6—"*The Burdens of Stable Judgment*"—poses the question, which arises even outside the context of trial proceedings, of whether the high probability of a proposition is enough for us to conclude that the proposition is true. I argue that high probability alone is not enough; we also need stability, i.e. a guarantee that future evidence will not lower or raise the current probability too dramatically.

Chapter 7—"*Four Ways a Reasonable Doubt Can Arise*"—turns to criminal trials and gives an account of the criminal standard of proof 'guilt beyond a reasonable doubt.' The account consists of four components: narrativity; high probability; evidential completeness; resiliency. To put it simply, a defedant's guilt is established beyond a reasonable doubt when the prosecutor has offered a well-specified narrative of the crime. The narrative, in addition, should be well supported by the available evidence and should have resisted the challenges from the defense.

Chapter 8—"*When is DNA Evidence Enough to Convict?*"—applies the framework of chapter 7 to DNA evidence. The main question of the chapter is whether DNA evidence alone is enough to justify a conviction. I argue that it is enough so long as DNA evidence contributes to supporting a well-specified narrative of the crime. I give an example of a DNA evidence case that satisfies this requirement, and an example of a case that does not. The chapter also contains a comparison between DNA evidence, eyewitness evidence, and fingerprint evidence.

Chapter 9—"*Looking back*"—draws some morals and conclusions.

Chapter 1

What Is Wrong with Statistical Evidence?

The evidence being presented in criminal trials is changing. Though jurors are still tremendously affected by traditional forms of evidence like a witness testifying 'it's him! I saw him!' or a fingerprint expert testifying that prints match, quantitative evidence is playing an ever increasing role in the courtroom. We have seen a steady rise in the use of statistics in criminal cases, especially as a result of the discovery of DNA fingerprinting in the 1980s. And relatedly, the phenomenon of "big data" has made it easier to find statistics for trial purposes. As early as thirty years ago, a commentator wrote that 'our criminal justice system is now at the threshold of an explosion in the presentation of mathematical testimony' (Jonakin, 1983, p. 369); this trend is even more marked today.[1]

While statistical and quantitive evidence increasingly gain momentum in the courtroom, traditional forms of evidence, such as eyewitness testimony and fingerprint evidence, are under attack. The testimony of a honest witness is the bedrock of the Anglo-

[1] A Lexis Nexis search through decisions by U.S. Appellate Courts over the last sixty years, in both civil and criminal cases, yields more than a thousand cases with 'statistical evidence' as one of the keywords. Half of these cases were decided in the last twenty years and one third of them were decided in the last decade.

American trial system, and fingerprint identification has long been considered the gold standard of criminal evidence. Yet, research in psychology and cognitive science has shown that these forms of evidence are often unreliable (Simons and Chabris, 1999), easily prone to manipulation (Loftus, 1996), and too much driven by subjective considerations and matters of context (Dror et al., 2006; Zabell, 2005). Concomitantly, the *Innocence Project* has shown that traditional forms of evidence, based on eyewitness testimony, confessions, and traditional forensic techniques, have been the cause of many wrongful convictions.[2] The attacks against traditional evidence—and more generally, against a traditional, non-mathematical way to understand criminal trials—stem also from a broader intellectual climate, in which quantification, precision, and mathematization are highly praised. These attacks cannot be dismissed nor ignored. However we think about them, they call for a reassessment of the role that traditional forms of evidence play in criminal trials.[3]

Statistical evidence has many of the features that traditional evidence lacks. It is more objective, precise, and scientific, and thus, it should be less likely to fall prey to the shortcomings that affect traditional forms of evidence. We may then feel tempted to seek comfort, so to say, in this new form of evidence. But we should make sure that, as we leave behind our blind faith in the old, we do not uncritically endorse the new. Some of us, in fact, feel uneasy about the use of statistics in criminal trials. While instructing

[2] As of May 2013, the Innocence Project found a total of 506 manifest cases of wrongful conviction; see *Innocence Project* at *http://www.innocenceproject.org/*.

[3] The judiciary and legal scholars have begun to explore ways to counter the dangers that traditional forms of evidence pose for the criminal justice system. Three strategies can be identified. The *first* consists in enhancing the training of police investigators, forensic experts, and legal professionals, so that the overall quality of testimonial and fingerprint evidence will improve, and at the same time, inadequate instances of them will be more easily screened off during pre-trial proceedings. The American Academy of Science has recently issued recommendations about forensic techniques; see NRC (2009). To improve the quality of eyewitness evidence, see the recommendations in Wells et al. (2006). The *second* strategy consists in making the admissibility rules for traditional forms of evidence more demanding. The Supreme Court of New Jersey in *New Jersey v. Henderson* (August 2011) has recently commented on the limits of eyewitness testimony and it has recommended more stringent admissibility standards. Finally, the *third* strategy consists in introducing rules of weight, e.g. a rule prohibiting that a criminal conviction be based on a single piece of uncorroborated eyewitness or fingerprint evidence; see Thompson (2008), Sangero and Halpert (2007), and Barzun (2007).

a jury about the use of statistical evidence, a British judge remarked: ‘we do not convict people in these courts on statistics. It would be a terrible day if that were so’ (Kadane, 2008, p. 409). Is the judge right? Would it be a terrible day if people were convicted on statistics? Why cannot statistics justify a conviction? A suggestion I shall explore in this chapter and throughout the dissertation is that statistics lack “specificity.” The idea is that statistics place the defendant in a group with others, and we do not want to be convicted because of what others did.

The plan for the chapter is as follows. In section 1.1, I formulate the question with which I will be concerned in this chapter. The question can be put as follows: is statistical evidence *alone* enough to justify a criminal conviction? In section 1.2, I offer two examples of how statistics are used in criminal cases. The examples involve DNA evidence and statistical correlations. In section 1.3, I turn to some criticisms of statistical evidence which have been formulated in the literature. I will mostly focus on the idea that what statistical evidence lacks is individualization and specificity. I will argue that the notion of specific evidence, albeit promising, does not hold water. In section 1.4, I sketch my own proposal. I will suggest that instead of speaking of specific evidence, we are better off speaking of specific narratives. In section 1.5, I discuss the controversial statement that all evidence is statistical or probabilistic; this will pave the way for the next chapter.

1.1 THE QUESTION(S)

Preliminarily, I should say what I mean by ’statistical evidence.’ It is difficult to offer a sharp definition, but we can identify a common pattern. Many of the statistics introduced in criminal trials are estimated frequencies. These can be *event frequencies* describing e.g. how many premature deaths occur among infants of middle-class families or they can be *feature frequencies* describing e.g. how often certain physical features or genetic markers occur in a select population. Such estimated frequencies are typically derived from actual frequencies or statistical data through the mediation of a statistical model. The estimates are then used to draw inferences bearing on guilt or innocence. All in all, we can identify three stages. The starting point is some quantitive data; next, from the quantitative data a statistical or probabilistic estimate is derived on the basis of a statistical model; finally, inferences bearing on guilt or innocence are drawn on the basis of the probabilistic and statistical estimates. As a matter of terminology, I will use 'statistics' and 'statistical evidence' almost interchangeably. The difference between the two is mostly in the emphasis: the word 'statistics' refers to the statistical data or the statistical estimates, while the expression 'statistical evidence' refers to the statistics insofar as they are presented in court as part of the evidence.

We may ask different questions about the use of statistics and statistical evidence in criminal trials. Some questions fall more squarely in the area of expertise of statisticians, probability theorists, and forensic scientists, such as:

(-) Have the statistics been properly calculated?

(-) How can we draw inferences from the statistics?

(-) Given the statistics, can a probability be assigned to a proposition that is relevant for deciding the case? How would one go about doing this?

These questions are best left to forensic scientists and statisticians. It is important, however, to realize that these questions—and especially the latter two—hardly have clear-cut answers. It is not the case that certain propositions can be given their own unique and correct probability assignment. Statisticians rely on assumptions and simplified models, and consequently, they might disagree on what assumptions are the most appropriate to use. In reflecting on the role of models and assumptions, some forensic scientists and statisticians who testified in court wrote:

> [We should not] overlook the subjective element in the choice of a probabilistic model and therefore the possibility of there being several models with very different predictions and perhaps different answers to different questions. The question as to how to use statistics . . . is not a question with a well-defined answer . . . many different approaches are possible, using very different models and with many different levels of sophistication. (Meester et al., 2007)

> [N]one of the proposed models is claimed to be, in some sense, 'right.' A [model] is a direct translation of one's subjective viewpoint toward an

> inference problem, which may be structured differently according to one's extent of background information and knowledge about domain properties. (Taroni et al., 2006, p. xv)

Other questions about the role of statistics in criminal trials are more strictly legal:

> (-) Is statistical evidence *relevant* given the rules of evidence?
>
> (-) Is statistical evidence *admissible* given the rules of evidence?
>
> (-) Is statistical evidence *enough to justify a conviction* given the criminal standard of proof?

In trial proceedings any evidentiary decision follows a three step process: for any piece of evidence, its relevance must be determined; then its admissibility; and finally its role in sustaining the final verdict. Against this background, it is natural to ask three (legal) questions about statistical evidence: whether it is relevant; admissible; or sufficient to justify a verdict of guilt.[4] I shall briefly discuss the relevance and admissibility of statistical evidence in an addendum to this chapter (see section 1.6). The question I am concerned with in this chapter and throughout the dissertation is the third one. It can can be put more precisely, as follows: *within our conception of a just criminal trial, is it acceptable to convict a defendant when the incriminating evidence consists merely of statistics*? To put it more succinctly, *are statistics alone enough to convict*?

To illustrate what the trouble with statistical evidence might be, let us begin with a hypothetical scenario. It is a simplified, artificial scenario, one which will hardly, if ever, resemble a real criminal case. Nonetheless, the scenario is useful to raise our question neatly; more realistic examples will follow shortly.[5] Here it is:

> *Prisoners*. A video recording shows that, in a prison yard, ninety nine out of one hundred prisoners present assaulted and killed the only guard on duty. In the recording, ninety nine prisoners appear to deliberately participate in

[4] There is also a further question, one in between the second and the third, namely whether statistical evidence alone is sufficient to make a prima facie case for the prosecutor. See See 94 Mich. App. 356, 288 N. W. 2d 426 (Mich. App. 1980).

[5] Redmayne (2008) makes a good case for considering hypothetical scenarios. Finkelstein and Levin (2001) offer a good survey of the uses of statistical evidence in the law.

> the killing; only one refrained. The faces are unrecognizable, and no other evidence is available. All prisoners escaped except one, Eschaton, who is captured and tried. Given the 99:100 statistical rate of participation in the killing, one might argue that Eschaton's guilt is 0.99 probable[6] and that Eschaton should be convicted because a high probability of guilt is enough to satisfy the criminal standard of proof. This argument, however, would be considered unacceptable by nearly everybody. [7]

Eschaton's guilt is highly probable, yet most of us have the gut reaction that we should not convict him on the basis of the statistics alone.[8] Not much would change if the odds were 999:1 or higher.[9] This suggests that there must be something wrong with a conviction that is based on purely statistical evidence even when the probability of guilt is very high. Convicting Eschaton would effectively be a gamble. We seem to lack any basis, after all, for thinking that Eschaton was not the one, lone innocent prisoner who did not participate in the killing of the guard.

[6] Here is the reasoning:

(S1) It is a fact that 99 out of 100 prisoners killed the guard;
(S2) Given (S1), Eschaton partook in the killing with 0.99 probability;
(S3) Given (S2), the probability that Eschaton is guilty is 0.99;

Premise (S1) rests on the assumption that the statistical rate of participation in the killing was properly estimated. Premise (S2) turns a statistical rate into a probability estimate attached to a proposition that is relevant to establish guilt. The proposition in question is 'Eschaton participated in the killing,' and a probability assignment of 0.99 stems naturally from the 99:100 rate of participation in the killing. Premise (S3) holds given the idealization that whoever participated in the killing is automatically guilty, and the scenario above seems to licence such idealization.

[7] This is a modification of a scenario in (Nesson, 1979, p. 1192).

[8] Wells (1992) tested people's reactions in scenarios similar to *Prisoners* and he demonstrated that most people feel uneasy in convicting *even when* the governing standard of proof is lower than 'beyond a reasonable doubt.'

[9] Or would it? Roth (2010) thinks that when the guilt probability is astronomically high our reluctance should disappear. She argues that this is precisely what happens with DNA evidence which can support an astronomically high probability of guilt. I disagree with Roth and I explain why in chapter 8.

1.2 Two examples

Some might complaint that—with the exception of the prison yard scenario—the evidence in a criminal case never consists of merely statistical evidence; there will alway be other items of evidence available. Although it is true that the evidence in a criminal case is never merely statistical, it can be *largely* statistical. What does that mean? The evidence is largely statistical whenever the prosecutor's case stands or falls depending on whether certain statistical estimates stand or fall. I will illustrate this point in detail in chapter 4, but let me offer here some representative examples.

1.2.1 DNA Evidence

Take DNA evidence whose probative value rests, to a great extent, on statistical and probabilistic estimates.[10] The basics of DNA evidence are relatively straightforward. When traces of blood, semen, saliva, skin tissue, etc. are found at the crime scene, laboratory analyses can create a DNA *profile* from the traces. A DNA profile is a codified representation of certain portions of the human genome which tend to be different across individuals.[11] Once a profile is created from the traces, it is compared against a suspect's blood, semen, saliva, skin tissue, etc. from which another DNA profile is created. The purpose of the comparison is to find a genetic *match* between the two profiles.[12] If a match is found, this would constitute evidence that the suspect is the source of the traces found at the crime scene.

[10] For a quick introduction to DNA evidence and its uses in the courtroom, see (Wasserman, 2008). For a more in-depth treatment, see (Kaye and Sensabaugh, 2000).

[11] On a chromosome we can individuate specific positions, called *loci*, which are "occupied" by a particular DNA sequence, called an *allele*. A DNA profile conveys information about the allele sequences at a select number of loci with high variability in alleles. In the United States, the Combined DNA Index System (CODIS) created by the FBI requires that a DNA profile consist of 13 select loci. For more information, see Kaye (2010b).

[12] DNA profiles created from different samples are never identical; they are *more* or *less* congruent. Typically, two profiles are said to match if they are sufficiently similar within a tolerance interval; see Kaye (1993). Further, forensic experts might be mistaken when they claim that two profiles match (e.g. because of contamination, switching of the samples, or because the two DNA profiles do not match at all). A reported match must not confused with an actual match; see (Thompson et al., 2003).

But a genetic match between two DNA profiles, in and of itself, is not particularly probative. What makes DNA evidence incredibly powerful is its statistical grounding. DNA profiles, albeit not unique, are highly discriminating because they each occur very rarely.[13] The rarity of a DNA profile is expressed by a statistical frequency, sometimes as astronomically small as 1 in 50 billion, representing the profile's expected frequency in a population.[14] The lower the statistical frequency, the more discriminating the profile and consequently the more probative the match.

In recent years, defendants have been convicted on the basis of evidence that consisted largely, or almost exclusively of DNA evidence. This is happening because of a new police practice. Besides traditional forms of crime investigation, police officers can now run a DNA profile created from the crime traces against a database of existing profiles (Roth, 2010; Devlin, 2007). When a matching profile is found in the database, the individual associated with it becomes the target of an investigation, and could later face trial. No doubt, in these cases the evidence against the defendant is largely statistical, in the sense that if the statistical calculations underlying DNA evidence turn out to be flawed, the prosecutor's entire case is undermined.

Some will think that DNA evidence does not represent a particularly problematic example of how statistics are used in criminal cases. On the contrary, it might represent an example of how good statistics can be. Some appellate courts in the United States, in fact, have declared that DNA evidence—given some particular circumstances—is enough to justify a conviction. For instance, New York highest court in People v. Rush (1995) wrote:

[13] Two individuals might share the same DNA profile; see (Weir, 2007) and (Saks and Koehler, 2008).

[14] A DNA profile coveys information about the DNA sequences (alleles) at select loci. Each allele has a certain frequency in a given population and such frequency is estimated by counting how many times the allele shows up in a database of DNA profiles. The bigger the database, the better the frequency estimate. The frequency of the entire profile is calculated by multiplying the frequencies of the single alleles. Consequently, the more alleles included in the profile, the lower the frequency. The multiplication of allele frequencies is justified on the assumption that each allele occurs independently of the others. This assumption is non-trivial and scientists debated it widely, though they have come to accept it; see (Kaye, 2010b). On a different note, Buckleton (2005b) invites us to take the astronomically low frequencies of DNA profiles with a grain of salt, because the statistical and genetic models from which they are derived cannot be tested empirically when frequencies are extremely low.

> [T]he testimony that there is a genetic match between the semen recovered from the victim of a rape and the blood of the defendant, a total stranger, and the statistical probability that anyone else was the source are 1 in 500 million is legally sufficient to support a guilty verdict.[15]

Furthermore, in the same decision New York highest court compared DNA evidence with eyewitness testimony. The court had no doubt that DNA evidence was better:

> [T]he perils of eyewitness identification testimony far exceed those presented by DNA expert testimony. Where the prosecutor is confronted with an irreconcilable conflict between eyewitness identification evidence and DNA identification evidence, it is likely to rely on the DNA evidence.[16]

Importantly, the *Rush* decision has been endorsed by a number of other appellate courts across the United States.[17] The decision is by no means uncontroversial, but it suggests that if statistical evidence is taken to be DNA evidence, statistics might not be as problematic as we thought at first.

1.2.2 Correlations

Courts hardly have a uniform and clear position on how to use statistical evidence in criminal cases. At times they are favourably inclined towards it (as in the case of DNA evidence), and other times they are very critical. To illustrate the critical cases, suppose one is charged with child abuse and is a father-figure for the abused child. Suppose, also, the prosecutor attempts to introduce base rate statistical evidence showing that 95 percent of child abusers are father-figures for the abused children. In absence of

[15] 630 N.Y.S.2d 631, 634. Note that the court talked of a probability of 1 in 500 million rather than a frequency of 1 in 500 million. It is better to speak of an estimated frequency or a probability because no actual counting was done; rather, the estimate 1 in 500 million is the result of genetic population models. Further, note that the court committed what is known as the prosecutor's fallacy because it talked about the probability that someone else was the source, whereas 1 in 500 million is the probability that someone who is not not source would be found with a matching DNA. The error here seems innocuous. For more details on this, see chapter 5.

[16] 630 N.Y.S.2d 631, 634.

[17] See e.g. U.S. v. Wright, 215 F.3d 1020 (9th Cir. 200); Roberson v. State, 16 S.W.3d 156 (Tex. App. 2000); State v. Abdelmalik, 273 S.W.3d 61 (Mo. App. 2008).

other evidence, US courts will dismiss the charges.[18] If the statistics express a likely correlation between possessing a certain feature (e.g. being a father-figure) and doing a certain action (e.g. committing violence on one's child), courts are very uneasy with using the statistical correlation to support a conviction.

Let us look at another example of a statistical correlation. The facts I will recount are from the case US v. Shonubi (1993). Charles Shonubi flew from Nigeria to the United States. He was found carrying 427.4 grams of heroin at JFK airport in New York. He was placed under arrest and went to trial. At sentencing, the federal district judge estimated the total amount of drugs Shonubi carried. The travel record showed that Shonubi made seven other trips between New York and Nigeria. The judge collected data from the US Custom Service about other Nigerian drug smugglers. A cursory look at the data showed that in the overwhelming majority of cases, drug smugglers carry more than 200 grams per trip. The expert witness for the prosecutor used a Monte Carlo simulation and concluded that there was a 99 percent probability that Shonubi carried at least 2090.2 grams of heroin during his seven other trips. Interestingly enough, although the judge at the trial level was willing to conclude that Shonubi carried roughly 2000 grams in total, the Appellate Court rejected the conclusion because it was based on mere statistical speculations (see chapter 4 for a fuller discussion). Note that here the matter is complicated because we are dealing with a sentencing hearing in which the governing standard is not beyond a reasonable doubt, but a lower standard, such as clear and convincing evidence. This complication is also what makes the case interesting. Even relative to a standard lower than the criminal standard of proof, some courts are unwilling to reach a conclusion on the basis of a statistical model.

One question that naturally emerges at this point is, what is the difference between the statistics about father-figures and drug smugglers and the statistics underlying DNA evidence? Is DNA evidence a better type of statistical evidence? If so, what makes it better? I do not think that statistical evidence is good or bad *per se*, nor do I think

[18] Stephen v. Wyoming. 774 P.2d 60 (Wyo. 1989). Washington v. Maule. 35 Wash. App. 287; 667 P. 2d 96 (1983). Washington v. Petrich. 101 wash. 2d 566, 683 P.2d 173, 180 (1984). Halle v. Arkansas. 15 Ark. App. 309, 692 S.W. 2d 769 (Ct. App. 1985). See (Koehler, 2002).

that some types of statistical evidence are better or worse. As I argue in section 1.4, what makes the difference is how the statistics are used and the types of inferences that are drawn from them. Before I turn to this topic, however, I will examine some of the arguments that have been offered for why statistics are problematic, and in particular, for why they cannot be enough to justify a conviction.

1.3 AGAINST STATISTICAL EVIDENCE

Existing arguments emphasize different problematic aspects of statistical evidence: prejudicial effects; lack of specificity and individualization; moral, socio-political, and procedural difficulties. I should emphasize that I am going to assume that the statistics are not grossly flawed nor are the result of a blatant statistical error. It goes without saying that if the statistics are flawed, they should never be enough to legally justify a conviction.

1.3.1 Prejudicial effects

The mismatch between how ordinary people actually reason with statistical information and how they should reason according to the norms of probability theory is well known (Tversky and Kahnemman, 1974, 1983; Kahneman and Tversky, 1982; Saks and Kidd, 1980; Koehler, 1996b; Gigerenzer et al., 1999; Kelman, 2011). Among others, a surprising result is the so-called birthday problem (Diaconis and Mosteller, 1989). What should the minimum number of people in a room be such that at least two of them share the same birthday with a probability of 50 percent? The intuitive thing to do here is to divide the number of days in a year by half. Yet, the correct answer, according to probability theory, is 23 people. Perhaps even more surprisingly is that the probability of finding two people with the same birthday raises to 99 percent if there are 57 people. This means that in a party with only 57 people, it is almost surely guaranteed that at least two people will share the same birthday!

The mismatch between our common sense intuitions, on one hand, and probability and statistics, on the other, can be used as a basis for an argument that jurors and judges should be better trained in probability and statistics. Yet, sometimes probability theory yields results that are so counterintuitive that even trained mathematicians are unwilling to believe them.[19] Understandably, then, the very same mismatch is often invoked for limiting the use of statistics in the courtroom. For instance, consider what happened when the Minnesota legislature in 1992 passed an act establishing that:

> In a civil or criminal trial or hearing, statistical population frequency evidence, based on genetic or blood test results, is admissible to demonstrate the fraction of the population that would have the same combination of genetic marks as was found in a specific human biological specimen. Minn. Stat. n. 634.26 (1992).

The Minnesota Supreme Court in State v. Bloom (1994) responded to the legislature very critically. It held that statistical population frequencies for DNA profiles cannot be presented to a jury.[20] This reaction may seem overly restrictive, and probably it is. Still, the Minnesota Supreme Court was worried that when jurors hear impressive statistics, they often uncritically interpret them as describing a very high probability of guilt.

The worry voiced by the Minnesota Supreme Court is not baseless. Suppose a juror hears that the defendant has a genetic profile that matches with the crime traces, and that the profile's frequency in the suspect population is, statistically speaking, one in ten million. The jurors might reason: 'Oh, this is a very low frequency, the defendant must be the perpetrator—he must be guilty!' There are at least two mistakes here. First, a frequency of one in 10 million means that, in a suspect population of 30 million, we expect to find three matching individuals. Second, no matter how impressive the statistics are, they do not automatically translate into a probability of guilt.

All in all, part of the problem with statistics and probability in criminal trials has to do with their prejudicial effects. As Lawrence Tribe, in his famous 1971 article on trial-by-mathematics, put it:

> Guided and perhaps *intimidated by the seeming inexorability of numbers*, induced by the persuasive force of formulas and the precision of decimal

[19] Here I have in mind the Monty Hall problem; see Devlin (2007).

[20] 516 N.W.2d ,159

> points to perceive themselves as performing a largely mechanical and automatic role, *few jurors ... could be relied upon to recall, let alone to perform, [their] humanizing function.* (Tribe, 1971b, emphasis mine)

The issue of the prejudicial effects of statistics and probability in the courtroom is certainly an important topic.[21] But the concerns about statistical evidence need not be merely psychological. Those who feel uncomfortable with the possibility that a conviction could rest on mere statistical evidence would feel uncomfortable *even when* the statistics had no prejudicial effect at all, *even when* jurors and judges were able to understand statistics and probability properly (Wells, 1992). It seems that the problem—if there is any—must run deeper than the question of prejudice.

1.3.2 Individualized evidence

Besides psychological worries, the criminal justice system has always been cautious with evidence based on probability and statistics. As a vocal expression of the judiciary's reservations, here is what a court wrote more than hundred years ago:

> Quantitative probability . . . is not proof, nor even probative evidence, of the proposition to be proved. That in one throw of dice, there is a quantitative probability, or greater chance, that a less number of spots than sixes will fall uppermost is *no evidence whatever that in a given throw such was the actual result*. Without something more, the actual result of the throw would still be utterly unknown. Day v. Boston & Maine R.R., 96 Me. 207, 217-218 (1902). Emphasis is mine.

The quotation is partly the expression of an unwarranted mistrust toward statistics and probability, a mistrust that has substantially declined in more recent years. Nevertheless, I find the quotation instructive. The court notes that the statistical evidence that a number less than six will come out is no evidence that *in this particular case* a number less than six has come out. This observation is vague, but it suggests that the incriminating evidence should be case-specific or individualized to the event under consideration, i.e.

[21] See, among others, (Thompson and Cole, 2007), (Koehler, 1993), (Koehler, 1996a).

this specific throw of a dice rather than another.[22] This naturally leads to the idea that the evidence should be specific to the defendant and the facts of the crime; it should not be about a group of individuals of which the defendant happens to be a member.

The notion of specific or individualized evidence has a great deal of intuitive appeal. In the prisoner case, for instance, the statistical evidence is not specific because it can apply to any prisoner, and not only to Eschaton. The probability of guilt, given the statistics, is high for any prisoner. This means that the statistics cannot sigle out Eschaton as (one of) the perpetrator(s) of the crime. The idea of specific evidence seems also to make sense from the point of view of forensic science. Forensic scientists, at least in the past, used to distinguish *class evidence* from *individualizing evidence*. Class evidence selects a class of individuals, e.g. the class of individuals having a certain blood type; individualizing evidence selects one, unique individual. Fingerprint evidence is—or better, was—commonly thought to be individualizing evidence.

At the same time, it has become clear that the distinction between class evidence and individualizing evidence is unsustainable, because no evidence is individualizing; no evidence can be uniquely specific to a particular defendant. Eyewitness evidence cannot be individualizing because two individuals can have practically identical faces, and an eyewitness is unlikely to discriminate between individuals with alike facial features. Neither are fingerprint evidence or DNA evidence instances of individualizing evidence. Fingerprints are not unique to individuals, nor are DNA profiles, or at least there is no scientific basis for thinking they are (Kaye, 2013; Saks and Koehler, 2008; Zabell, 2005). Specific evidence—at least in so far as it is grounded in the individualizing/class evidence distinction—represents an unreachable ideal. It is what is commonly referred to as the "smoking gun"—evidence that can tell us, with absolute certainly, who did the deed. But the evidence in criminal trials is hardly of this sort.

Against this background, the following quotation, albeit written some thirty years ago, is still timely:

[22] The role of tense seems also important here. The statistical chances can be used to predict that a certain outcome will result. Instead, traditional forms of evidence exists only after an event has occurred; they cannot be used to predict, but only to retrodict.

> Perhaps the most serious error is an epistemological one: the assumption that case-specific information is really qualitatively different from base-rate information . . . And, indeed, it seems obvious that background base-rate information is about other cases while particularistic information is about this case. . . . Much of the testimony that is commonly thought of as particularistic only seems so. . . . This includes eyewitness identification . . . , fingerprints . . . , and anything else we could name. . . . All identification techniques place the identified object in a class with others . . . There is little, if any, pinpointed, one-person-only evidence in this world. (Saks and Kidd, 1980, p. 151).

Despite the seemingly unobjectionable fact that any form of evidence—DNA, fingerprint, eyewitness evidence—tends to 'place the identified object in a class with others,' courts sometimes operate on the assumption that the notion of specific evidence makes sense. Judge Cardamone of the Second Circuit Court of Appeal in the (already mentioned) drug trafficking case US. v. Shonubi (1993) writes:

> Case law uniformly requires specific evidence—e.g. drug records, admissions or live testimony—to calculate drug quantities. 998 F.2d 84, 90.

The notion of specific-evidence, albeit intuitive, is highly contentious, and the district court Judge Weinstein, who decided the case at the trial level, disagreed vehemently with the Second Circuit:

> The new vaguely defined classification of "specific evidence" relied upon by the court of appeals in Shonubi . . . is not only unauthorized by controlling case law and the federal Rules of Evidence, it runs counter to our modern theory of forensic evidence. The unique Second Circuit rule represents a retrogressive step. 962 F. Supp. 370, 375 (E.D.N.Y. 1997).[23]

This suffices to make the point that, though specific evidence is an initially appealing notion, it is very contentious and hard to grasp. I think, however, that the notion of specificity is a promising one and should be investigated further. I will suggest that specificity is better understood as a feature that applies to narratives, stories or reconstructions of the crime as a whole, rather than to isolated pieces of evidence. The rough

[23] See e.g. Tillers (1997) for a discussion of the case.

idea is that stories or narratives as a whole, not pieces of evidence, are specific whenever they describe the crime in a complete, exhaustive, and satisfactory manner. But before I expound on my narrative-based approach to specificity, I shall examine some other criticisms of statistical evidence we find in the literature.

1.3.3 Moral, political, and procedural considerations

So far I have examined psychological and epistemic considerations regarding what might be deficient about statistical evidence. But discussions of statistical evidence and its shortcomings have also focused on other types of considerations. Three examples of what scholars have suggested should suffice here. First, Pundik (2009) makes a point that rests on *moral* considerations. He observes that a criminal trial attributes culpability to the convicted defendant, on the assumption that the defendant is a unique individual and a free agent. He argues that if an attribution of culpability rested on statistical evidence—and in particular, on evidence about the frequency of certain behavioural features in a group—this would be inconsistent with the assumption that defendants are unique and free agents.[24]

Second, Enoch et al. (2012) are more interested in the *socio-political* dimension. They ask us to imagine what would happen if convictions were based on statistical evidence. The societal perception would be that defendants are convicted, not on the basis what they did or did not do, but on the basis of some group-based statistical regularities. As soon as citizens realize that it does not matter what they do or do not do as individuals, this would undermine the deterrence function of convictions.[25]

Finally, Stein (2005) offers considerations that we may call *procedural*, because they have to do with trial proceedings and practices. He notes that statistical evidence is at odds with cross-examination, a well-established practice which affords defendants

[24] On the moral front, Tribe (1971b) thinks that statistical evidence dehumanizes the trial process, and in particular, Zuckerman (1986) and Wasserman (1991) think that it undermines the defendant's individuality and freedom.

[25] On the socio-political front, Sanchirico (2001) complains that statistical evidence creates distortional incentives for future criminals and Nesson (1979) thinks that it renders judicial verdicts too easily subject to social scrutiny.

an opportunity to challenge the prosecutor's case. Statistical evidence is at odds with cross-examination, Stein thinks, because the latter works best when the incriminating evidence is case-specific, and presumably statistics are not case-specific.

Importantly, the moral, socio-political, and procedural considerations I've just outlined do not stand on their own. They all rest on an implicit conceptualization of certain features that statistical evidence would possess or lack. For example, Pundik indicates that statistical evidence is about a group and not the individual, so that using it for a conviction would be in tension with the assumption that attributions of culpability apply to defendants as unique and free agents. Enoch et al. explicitly rely on the idea that statistical evidence does not "track" what individual agents do or do not do, and hence convictions based on it woud fail to have any deterrence effect. Finally, Stein hints at the fact that statistical evidence is not case-specific and thus it would be at odds with cross-examination. All in all, moral, political, and procedural criticisms of statistical evidence rely on some prior characterization of statistical evidence, e.g. as evidence concerning only the behavioral features of a group and not of the individual (Pundik), as evidence lacking the ability to "track" a person's actions (Enoch), or as evidence lacking case-specificity (Stein).

1.3.4 Different analyses

No matter what we think about statistical evidence in criminal trials, my brief examination of the literature showed that we can pursue different analyses:

> *psychological*: how the fact-finders reason and how they react to statistical evidence (e.g. prejudicial effects);
>
> *evidential*: features statistical evidence lacks (e.g. specificity, causal connection);
>
> *moral*: whether statistical evidence is at odds with some moral values in trial proceedings (e.g. defendant's autonomy and freedom);
>
> *socio-political*: whether statistical evidence is at odds with certain social goals (e.g. deterrence); and

procedural: whether statistical evidence fits with well-established trial procedures (e.g. cross-examination).

In the dissertation, I will focus primarily on the evidential and the procedural aspects. A programmatic remark I can make at the outset is that we will be in a better position to assess the role of statistical evidence once we weave together evidential and procedural considerations. But in order to appreciate how the two can go together, it is now time to outline the narrative-based approach to specificity which I have gestured at earlier. This approach rests on the idea that there is another way to look at statistical evidence, namely in terms of the inferences and conclusions we can draw from the evidence. To see what this means, we should go back to the notion of specific/individualized evidence and then gradually transition to the notion of a specific narrative.

1.4 Narrative specificity

The idea of individualized/specific evidence is that some items of evidence mark out an individual, his actions and whereabouts. Eyewitness identification or fingerprint identification would be examples of individualized evidence. In contrast, so the idea of individualized evidence goes, other forms of evidence can only place an individual in a group with others. Statistical evidence, then, would be a form of non-individualized evidence. Can we make sense of the distinction? I've already noted that two or more people might share the same fingerprint profile; two or more people might share the same facial features. If, as I argue in chapter 8, fingerprint identifications rest on matching prints and if eyewitness identifications rest on matching faces, both items of evidence would fail to single out one individual uniquely. They would both place an individual in a group with others, and little would be left of the intuitive idea of individualized evidence.

1.4.1 Evidence, inferences, and narratives

Fingerprint, eyewitness, and DNA evidence seem to a follow a common pattern. In all the three cases, an identification is carried out through the mediation of a non-unique identifying feature, i.e. a fingerprint profile, a face, or a genetic profile. As far as the intuitive notion of individualized evidence is concerned, no difference can found between these three types of evidence. (See chapter 8 for a defense of this point.) And yet, one might insist that fingerprint and eyewitness evidence are still quite different from statistical correlations. How so? Suppose I say: 'people fitting profile π are statistically more likely than people fitting profile not-π to do action α; you fit profile π, so you are more likely to do α.' We are often uncomfortable with this profile-based reasoning, especially if we were to replace π with 'father-figure' and α with 'rape your child.' It certainly does not seem right to say in a court of law 'the accused is a father-figure, so he is more likely to have raped his child.' Still, it seems true that violence against minors is often perpetrated by father-figures. In contrast, suppose I say: 'your fingerprints are of type τ, and the fingerprints at the crime scene are of type τ as well, so you probably visited the crime scene because profile τ is statistically infrequent.' This form of reasoning is perfectly acceptable, especially in a court of law. But note that both the fingerprint profile τ and the racial profile π pick out a number of people; neither of the two picks out one individual uniquely. What is the difference between the two cases, then?

I can think of three differences. First, the father-figure profile picks out a larger class of people than the fingerprint profile. This might be true, but we can think of other parental or behavioural profiles that pick out a smaller class of people (see e.g. the George Metesky example later). Second, the father-figure profile has more to do with a sociological datum, while a fingerprint profile is more firmly grounded in a physical trait of the individual. But if both statistics can pass the scrutiny of a well-qualified statistician, I do see why this second difference should matter. Third, the conclusions we draw from the statistics are different: 'the defendant did (or is likely to have done) action α' *versus* 'the defendant visited (or is likely to have visited) the crime scene'. This difference, I think, is crucial; it amounts to a difference in the *inferential role* of the statistics in the two cases. Let me explain.

I want to suggest that there is no intrinsic difference between the evidence in the two cases. The difference is in the inferences, not in the statistics themselves. But what, exactly, is the difference in the inferences? There does not seem any principled difference between drawing the conclusion that the defendant did action α and drawing the conclusion that the defendant visited the crime scene. Doing action α and visiting the crime scene, after all, are both actions of some sort. What is the difference, then? The difference, I hold, is in the degree of spatiotemporal specificity of the two conclusions. In drawing the inference that the defendant visited the crime scene, a specific time and place will presumably be given. In drawing the inference that the defendant did action α, instead, no detail about a time and a place might be given. So the difference in specificity is in the conclusion being drawn, not in the evidence itself. (At best, the evidence itself might be more or less specific depending on whether it supports a more or less specific conclusion.)

A new question now aries. Why does the difference in specificity (of the conclusions being drawn) matter? This can be appreciated only if we view trial proceeding as a process in which the prosecutor is expected to offer a coherent and reasonably well-specified narrative of the crime. Constructing a narrative, after all, is precisely a way of drawing inferences from a body of evidence and weaving those inferences together. In chapter 7 I will develop the notion of a crime narrative and discuss its role in criminal trial. Let me now only sketch a few ideas.

In establishing a defendant's *guilt* for a crime (murder, rape, theft, etc.), the prosecutor should prove a number of *factual propositions* from which guilt follows in accordance with the substantive law governing the case.[26] To prove the factual propositions of interest, the prosecutor typically advances a *well-specified narrative* (a story, a theory) of the crime which describes what happened in a coherent way.[27] The narrative should be supported by the *evidence* available,[28] and the factual propositions that needs to be

[26] E.g. according to the common law, to prove murder, the prosecutor is expected to prove the *actus reus* as well as the *mens rea*. It is customary to distinguish matters of fact (e.g. whether the defendant intentionally killed the victim with a gunshot) from matters of law (e.g. whether intentionally killing with a gunshot is a second or first-degree murder). The distinction might be more nuanced; see e.g. (Hruschka, 1965), (Allen and Pardo, 2003), (Friedman, 1992), and (Taruffo, 2009).

[27] On the role of narratives in the reconstruction of crimes, see (Anderson et al., 2005).

[28] At its simplest, the narrative should be highly probable on the evidence. But there might be additional requirements that the relation of evidential support should satisfy. Fitelson (2006) offers an excellent discussion of many probability-based accounts of evidential support.

established should follow from the narrative as matter of logic and common sense.[29] The narrative should be well-specified in the sense that it should offer a sufficiently specific and detailed reconstruction of what happened.[30] The relation of evidence, narrative, factual propositions, and guilt can be visualized as follows:

$$\begin{tikzcd}
\textit{narrative} \arrow[r] & \textit{factual propositions} \arrow[r] & \textit{guilt} \\
\textit{evidence} \arrow[u, dashed] & &
\end{tikzcd}$$

1.4.2 Specific narratives, not specific evidence

There is a fluid relation between evidence, inferences, and crime narratives, as will be suggested in chapter 7. Consequently, talks of specific evidence are artificial insofar as it is artificial to isolate items of evidence and attribute good or bad epistemic qualities to them, as if they existed independently *qua* things in themselves. There are no clear-cut criteria to individuate items of evidence. This is especially apparent in the case of DNA evidence. What is DNA evidence? Is it just a genetic match and a profile's frequency? What about the shape and quantity of the genetic material? What about the lab error rates? It is difficult to demarcate precisely what is part of DNA evidence and what is not.

To be sure, there is something about statistical evidence that might classify it as lacking individualization. Let me explain. If the prosecutor is expected to offer a well-specified narrative of the crime, statistical evidence *alone* seems typically incapable to support such a specific narrative. For instance, statistical evidence may tell us that people in a certain category are likely to steal in a clothing shop during a certain time of

[29] E.g. suppose the narrative asserts that the defendant stabbed the victim's with a knife and that the victim's death occurred immediately thereafter (as per medical doctor's testimony); the factual proposition that the defendant caused the victim's death follows from such narrative as a matter of shared knowledge and logic.

[30] As I argue in chapter 7, a narrative is well-specified if it gives an account of the co-occurrence of the *mens rea* and the *actus reus*. This means, for example, that the narrative should describe what motivated the accused to perform certain actions, what goals drove the accused, and what consequences resulted from those actions. This is what I called an *event-narrative*. Another feature that makes a narrative well-specified is that it can accomodate (or "explain") the presence of certain crime traces (physical, cognitive, or digital). This is what I called an *evidence-narrative*.

the day, but it will not give us a more specific time and place. It would be surprising that a piece of statistical evidence could tell us that certain people steal—or should I say: stole? will steal?—at Barbara Bui boutique in New York City on January 20th, 2006, at 4PM. Such a degree of spatiotemporal detail is typically unattainable for statistical evidence. In contrast, it should not be surprising if an eyewitness reports having seen, say, Jack steal in a precise place at a precise time. The level of narrative specificity that statistical evidence allows, then, would be lower than that of, say, eyewitness evidence.

We might say, then, that statistical evidence typically support less specific narratives than traditional forms of evidence do. But this suggestion is correct only in a very limited sense. The matter is complicated by the fact that DNA evidence, a form of statistical evidence, can support more or less specific narratives (see discussion in chapter 8). Moreover, there have been cases in which, maybe surprisingly, predictions based on statistical data turned out to be extremely detailed. George Metesky, known as the Mad Bomber, terrorized New York in the forties and fifties by placing explosives in various public places in the city. While investigating the bombing and trying to identify the perpetrator, Dr. Brussels, a criminologist and psychiatrist, put together a profile of the bomber by means of statistical data. The profile contained an incredible degree of detail: male, forty to fifty years old, neat and tidy, a loner, a Catholic, and much more. Dr. Brussels also predicted that when the bomber was apprehended by the police, he would be wearing a double-breasted suit, buttoned. Well, when the police found Metesky, not only did he match the profile, but also, he was wearing a double-breasted suit, buttoned! (Schauer, 2003).

1.4.3 Two advantages

All in all, my working hypothesis is that the enterprise of finding an inherent feature of statistical evidence which would make it deficient, irrespective of its contribution within a larger crime narrative, is hopeless. Statistical evidence, as any other evidence, has an inferential role (e.g. it supports a crime narrative); its probative value has to be assessed relative to its inferential role and on a case-by-case basis. I will now conclude this section by mentioning two advantages of a narrative-based approach to specificity.

One advantage of a narrative-based approach to specificity is that it fits well with the dialectical nature of criminal trials. Earlier in 1.3.3 we have seen Stein's criticism of statistical evidence. His view is that statistical evidence is at odds with the process of cross-examination because it is not case-specific. This observation is intuitive enough. It would be odd to cross-examine a statistician who is presenting some statistics in court. Since the statistician did not witness the crime, she would be unable to speak about what happened; she could only offer some "statistical speculations." And yet, there are plenty of ways to cross-examine a statistician, e.g. by scrutinizing the methods and procedures that were used to collect the statistics. What we cannot do, however, is to cross-examine the statistician *about what happened*; of the events of the crime, the statistician has no first-hand knowledge.

I think we can make better sense of Stein's observation that statistical evidence is at odds with the process of cross-examination if we take specificity to be a feature of crime narratives. Indeed, a close correlation exists between the degree of specificity of an incriminating narrative and its susceptibility to the challenges which the defense may raise. (As a matter of terminology, I prefer to speak more generally of 'challenges' rather than 'cross-examination.') To illustrate, suppose the prosecutor's narrative in a homicide case is relatively under-specified: it does not describe the weapon used in the crime nor does it say how the victim was killed. In order to mount a challenge against the under-specified narrative, the defense will have to target all the possibilities that have been left open. In this case, if the defense offerred evidence that e.g. the killing did not happen through strangulation, the prosecutor could easily respond that he did not specify how the killing occurred, and that if the killing was not by strangulation, it was by some other, unspecified procedure. The challenges—to be at least *prima facie* successful—will have to address all possible ways the killing could have occurred, such as poisoning, strangulation, stabbing, and many others. All in all, a quite extensive body of arguments would be required to challenge—*prima facie*, let alone successfully—the narrative in question. In contrast, if the prosecutor's narrative specifically said that the perpetrator poisoned the victim, the defense would only need to formulate a challenge that was pertinent to poisoning; it would not need to consider other ways the perpetrator could have killed the victim. In short, the following correlation exists: a more specific narrative is more susceptible to challenges because it says more, it commits itself to more propositions than a less specific narrative which leaves many issues and questions open. So an under-specified narrative is at odds with cross-examination because it commits itself to a limited number of assertions, and in so doing, it keeps its susceptibility to challenges artificially low. (More on this in chapter 7.)

A second advantage of a narrative-based approach is that it can distinguish good from bad uses of statistical evidence. To illustrate, recall the contrast between DNA evidence and other types of statistical evidence (see section 1.2). Suppose you are charged with child abuse and you are a father-figure for the abused child. The prosecutor attempts to introduce base rate statistical evidence showing that 95 percent of child abusers are father-figures for the abused children. In absence of other evidence, U.S. courts will dismiss the charges against you.[31] On the other hand, suppose you are tried for rape and your DNA profile matches with the traces found at the crime scene. The prosecutor attempts to introduce statistics showing how improbable it is that a random person would have a DNA profile that matches with the traces found on the crime scene. Many U.S. courts will find the statistics relevant and admissible (Koehler, 2002). They might even be enough for a conviction (see e.g. the *Rush* decision.) What makes DNA evidence a better type of statistical evidence? Is it simply that the probability of guilt on DNA evidence is higher or is there more to be said? My answer is that DNA is not necessarily a better type of evidence. *It depends on how DNA evidence is used.* So long as DNA evidence is used to support a sufficiently specific narrative of the crime, DNA evidence should raise no peculiar worries. By the same token, the father-figure statistics should raise concerns whenever they are used for offering an under-specified narrative of the crime. Once again, we are better off shifting our focus from the evidence (statistical or otherwise) to the crime narratives that are being offered by the prosecutor at trial. (More on this in chapter 8.)

[31] Stephen v. Wyoming. 774 P.2d 60 (Wyo. 1989). Washington v. Maule. 35 Wash. App. 287; 667 P. 2d 96 (1983). Washington v. Petrich. 101 wash. 2d 566, 683 P.2d 173, 180 (1984). Halle v. Arkansas. 15 Ark. App. 309, 692 S.W. 2d 769 (Ct. App. 1985).

1.5 Is all evidence statistical?

It is now time to turn to those who think that statistical evidence does not pose any peculiar problem for criminal trials. I am inclined to agree. As suggested earlier, the problem is not with the statistics *per se*, but with how they are used. Often we hear a statement along the lines that all evidence is statistical or probabilistic. Judge Richard Posner, a vocal advocate of the law and economics movement, writes in U.S. v. John Veysey (2003):

> All evidence is probabilistic—statistical evidence merely explicitly so. Statistical evidence is merely probabilistic evidence coded in numbers rather than words. . . . An eyewitness does not usurp the jury's function if he testifies that . . . he is "99 percent" positive. The signicant question would be the accuracy of the estimate.

The statement that all evidence is statistical is clearly false if it means that all evidence comes in a numerical form. Fingerprint and eyewitness evidence do not come with numbers attached. Yet, the statement that all evidence is statistical is less obviously false if it means that the probative value of any type of evidence can be interpreted in a quantitive way. We can spell out this idea more precisely as follows: any form of fallible evidence can establish guilt with a certain probability, a probability which will (almost) always be short of one hundred percent. So, for any body of evidence presented at trial we can write something like '$P(G|E) < 1$,' i.e. the probability of guilt given the evidence is less than one hundred percent. This conclusion can be taken to suggest that there should be no peculiar problem with statistical evidence. If all evidence is probabilistic in the way just described, it should not matter any longer whether the evidence is overtly statistical or not. What should matter, instead, is whether the evidence supports the conclusion of guilt with a sufficiently high probability *regardless of whether the evidence is statistical or not.*

But is it true that all that matters to justify a criminal conviction is the probability of guilt on the evidence? This probabilistic way of thinking about statistical evidence in criminal trials, and about evidence more generally, raises another question. It can be formulated as follows: *is a conviction legally justified on the sole ground that the probability of guilt—based on the available evidence, statistical or not—reaches a sufficiently high value?* I examine this question in the next chapter.

1.6 ADDENDUM: Relevance and admissibility

In this addendum, I examine whether statistical evidence is deemed irrelevant or inadmissible. I argue that in this respect US law is not particularly restrictive toward statistical evidence.

Relevance. To understand the notion of relevance, an obvious starting point is rule 401 of the Federal Rules of Evidence (F.R.E., hereafter), which defines relevant evidence as follows:

> Relevant evidence means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence. (F.R.E. 401)

It is customary to distinguish two dimensions operating in this definition: materiality and probative value (Fisher, 2008; Méndez, 2008). A piece of evidence E is probative of a fact F, provided E makes F more (or less) probable than it would be in absence of E. On the other hand, a fact F is material for the action or crime C under examination if it is of consequence to the determination of C. Whether or not a fact is of consequence to the determination of C depends on the substantive law which determines what facts should be proven in order to establish the crime.[32] By combining materiality and probative value, it follows that a piece of evidence E is relavant to crime C if and only if it is probative of F, where F is material for C.

To determine whether statistical evidence is relevant, one should consider materiality as well as probative value. There may be cases in which the substantive law explicitly requires that statistics be introduced for showing the existence of a certain disputed fact. This occurs, for instance, in discrimination controversies in which the plaintiff has to

[32] This is a simplification, however. More precisely, a fact can be material if (a) it is an ultimate issue (one that the governing law says should be proven to establish the crime charged), or (b) it is an intermediary fact, or (c) it is an evidentiary fact. The governing law establishes materiality only for ultimate issues, not for the intermediary and evidentiary facts.

prove a statistical correlation between sex and differences in pay.[33] If so, statistical evidence will be *ipso facto* relevant. More often than not, however, the law will not explicitly require the introduction of statistical evidence, so that the question of relevance will depend on the probative value of the statistics.

Consistently with F.R.E. 401, evidence E has probative value whenever it increases the probability of the material fact F, i.e. $P(F|E) > P(F|\neg E)$. On this characterization, does statistical evidence have probative value? Take the prison yard scenario from section 1.1. Suppose we did not know that 99 out of 100 prisoners participated in the killing, but only that some did, maybe two or three; then, the guilt probability of Eschaton would be quite low. But as soon as we learn that 99 prisoners killed the guard, his probability of guilt reaches 0.99. The statistics are therefore relevant here. And, I suspect, we would arrive at a similar conclusion as to relevance in most cases involving statistical evidence, provided the defendant's probability of guilt on the basis of the statistics is high.[34]

Admissibility. And now I turn to the admissibility of statistical evidence. The question of admissibility is a strictly legal one, and it depends on the law of the particular jurisdiction. For instance, in federal cases, F.R.E. 402 defines admissibility as follows:

> All relevant evidence is admissible, except as otherwise provided by the Constitution of the United States, by Act of Congress, by these rules, or by other rules prescribed by the Supreme Court pursuant to statutory authority. Evidence which is not relevant is not admissible. (F.R.E. 402)

To determine admissibility, it is sufficient to determine whether a piece of evidence is relevant, and whether it is deemed inadmissible by some exclusionary rule found in

[33] See the US Equal Pay Act, section 206(d)(1).

[34] Similarly, probative value can be captured by using likelihoods, i.e. E is probative of F whenever $P(E|F) > P(E|\neg F)$. See Lempert (1977) and (Royall, 1997). Bayes' rule ensures that if a piece of evidence E raises the probability of F, then $P(E|F) > P(E|\neg F)$. So the statistics in *Prisoners* would be relevant on this characterization as well. Some authors suggested that evidence E has probative value whenever it is causally connected to fact F or it explains F, given an account of causality and explanation; see (Achinstein, 1978). On this characterization, the statistics in the prison yard scenario do not count as relevant because they do not explain that Eschaton killed the guard, but traces evidence such as fingerprints would.

the F.R.E., the U.S. Constitution, an Act of congress, or rules prescribed by the U.S. Supreme Court.

In the F.R.E. there are at least two exclusionary rules that could block the admissibility of statistical evidence. One is F.R.E. 403:

> Although relevant, evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury, or by considerations of undue delay, waste of time, or needless presentation of cumulative evidence. (F.R.E. 403)

Rule 403 requires the judge to perform a balancing test between the probative value of the evidence and certain negative or detrimental effects (prejudice, confusions, waste of time, etc.) that are more or less likely to occur if the contested piece of evidence is introduced. Rule 403 is a curious rule, and its standard of application is far from being well defined.[35] One way to interpret the rule is to say that it requires the judge to compare and balance a purely epistemic value—i.e. the probative value of a piece of evidence—with other non-epistemic values or dis-values, such as prejudice, confusion, and waste of time.[36] The rule states that probative value is not always enough to guarantee admissibility, and hence a balancing test between epistemic and non-epistemic dimensions must be performed by the judge. The balancing-test of the two dimensions assumes that they can be compared using some common metric.[37] In balancing epistemic and non-epistemic dimensions, statistical evidence may turn out to be inadmissible. If, for instance, the statistical correlation is incredibly high or incredibly low, the jury may be

[35] For instance, what are the items that are intended to be balanced? The degree of probative value with the likelihood of prejudice, confusions, waste of time, etc.? Or, the degree of probative value with the degree of "detrimental seriousness" associated with prejudice, confusion, waste of time? Or does the balancing include all three variables?

[36] To be sure, prejudice and confusion might be epistemic (dis)values.

[37] Some have doubted that any balancing test between the two dimensions is possible because it would be like comparing apples and oranges, as it were; see Taruffo (2009). On the other hand, balancing tests are ubiquitous in the law, especially to reconcile conflicting values and goals. For instance, in 4th Amendment law about (un)reasonable searches and seizures two competing goals must be reconciled. One goal is the state's need to collect evidence and the other is the citizen's right not to have their privacy rights violated. The Supreme Court solution, in some cases, is to adopt a balancing-test. It is an interesting question whether the "balancing strategy" is the only, or the most appropriate strategy to undertake in case of a value conflict.

impressed by it and thus uncritically decide for a conviction without carefully examining the significance of the numerical quantification. This is a prejudicial effect that could substantially outweigh the probative value of the evidence according to rule 403.[38]

The other exclusionary rule contained in the F.R.E. that may affect the admissibility of statistical evidence regulates character evidence. In particular, rule 404(b) establishes:

> Evidence of other crimes, wrongs, or acts is not admissible to prove the character of a person in order to show action in conformity therewith. (F.R.E., 404(b))

Character evidence concerns the previous conduct of the defendant, not the specific conduct under examination in the trial.[39] So long as statistical evidence concerns the defendant's behavioural tendencies, it will count as a type of character evidence and thus it would be excluded. In the prison yard scenario, the statistics concern the acts of other individuals, or in the case of DNA evidence, the statistics concern the frequency of a DNA profile in a population. Hence, rule 404(b) against character evidence does not openly declare statistical evidence inadmissible.

From this brief overview, the rules governing evidence and procedure appear to be open to the introduction of statistical evidence; no rule explicitly declares it irrelevant or inadmissible per se.

1.7 ADDENDUM: CAUSALITY AND LUCK

In this addendum, I examine an interesting account of individualized evidence by the moral philosopher Judith Thomson (1986). She thinks that an item of evidence counts as individualized provided it is casually connected, in the appropriate way, to the event or state of affairs that the evidence is purported to establish. (This characterization is different from the intuitive notion of individualized evidence I have discussed in this

[38] To gain further insights into the applicability of rule 403 to statistical evidence, it would be interesting to review pertinent court's decisions. The cases cited in Koehler (2002) are a good starting point.

[39] Interestingly, some have recently argued for the relevance and admissibility of character evidence; see (Redmayne, 2002).

chapter, so it deserves further consideration.) On her account, an eyewitness identification would be individualized evidence because, presumably, the witness was, at some point in time, causally connected through her sensory apparatus with the person being identified as the perpetrator. Incidentally, it is unclear whether DNA evidence would count as individualized for Thomson. The DNA statistics are not causally connected, but the genetic traces certainly are. This unclarity, it seems to me, stems from Thomson's implicit assumption that items of evidence can be isolated and demarcated; on this point, see section 1.4.

For Thomson, no criminal conviction can rest on evidence that is non-individualized, i.e. non-causally connected with the event of the crime. Why should the evidence be causally connected with the event of the crime? Thomson's answer is that otherwise it would be a matter of luck whether a verdict ends up being right. We are here dealing with epistemic luck, not moral luck. Epistemic luck is the type of luck that derives from being accidentally right about a certain factual question e.g. whether the accused is guilty or innocent. The case Thomson has in mind is one in which a guilty verdict rests on the mere statistical chances that a defendant is guilty. Think about the prison yard scenario which we have encountered earlier in section 1.1. In that scenario, the defendant Eschaton is on trial because the statistics-based probability that he killed the prison guard equals 99:100. Intuitively, it would be a matter of (epistemic) luck if the jury convicts Eschaton, and Eschaton is in fact guilty. Thomson thinks that, in contrast, if the jury relied on a causally-connected item of evidence pointing toward Eschaton's guilt, the jury would be *ipso facto* shielded from epistemic luck. Why would that be so? This is by no means obvious. Suppose that, on the basis of witness testimony, a defendant is convicted, and that the defendant is in fact guilty. Now, the verdict might still be correct as a matter of luck. After all, the witness could have been wrong; he could have misidentified the defendant; he could have been lying. If none of thee possibilities in fact materialized, isn't that still a matter of luck? And how are we to understand epistemic luck in the first place?

But let us grant Thomson that *whenever* the evidence is causally connected with the event of the crime, the causal connection frees the fact-finders from epistemic luck. A problem still remains. No onlooker can tell beforehand whether an item of evidence is *de facto* causally connected. Eyewitness evidence is causally connected in some instances, but in many other cases it is not. Suppose the witness on the stand is lying or has misperceived the event of the crime. These would be cases in which the putative causal connection is absent or failed in some relevant way. So, how useful is it to have an account of individualized evidence which relies on a feature—i.e. the causal connection with the event of the crime—that cannot be grasped by the fact-finders? Thomson might insist that this is precisely a feature of her account: it is an *externalist* account of evidence. Yet, unless the notion of a causal connection can be operationalized in some way, I do not see how it can be of service to understanding criminal trials.

These difficulties notwithstanding, Thomson's account of individualized makes significant progress. As noted in section 1.4, my strategy consists in shifting from 'evidence specificity' to 'narrative specificity.' In this spirit, Thomson requirement of a causal connection between the evidence and the event of the crime can be understood as a requirement that a *causal narrative* be presented at trial. This narrative would show how (some of) the evidence presented at trial is causally connected with the facts of the crime. In this way the causality requirement would no longer be a metaphysical, ungraspable feature of the evidence. (This suggestion is developed more fully in chapter 7.)

I also think that Thomson's observation that a verdict cannot be right as a matter of epistemic luck is essentially correct. The problem with this is that we do not have a worked out account of the notion of luck in criminal trials.[40] Here is how I propose to make sense of this notion. For one, the elimination of epistemic luck cannot amount to the full-fledged infallibility of the trial system; some errors will be inevitable no matter what. Note, however, that one of the goals of a well-functioning, albeit still fallible, trial system is to reduce errors *as much as possible*. So it is plausible to think that a verdict would be free from epistemic luck whenever it is the outcome of a trial process that strived to reduce errors as much as possible; conversely, a verdict would be subject to

[40] Pritchard (2005) offers a general account of the notion of luck. His account is interesting on its own but it is not immediately applicable to criminal trials.

epistemic luck whenever it is the outcome of a trial process that failed to reduce errors as much as possible. This account of luck in criminal trials is developed more fully in chapter 2 (see, in particular, the discussion of Thomson's account in the addendum to that chapter).

Chapter 2

Is High Probability Enough to Convict?

The marriage between probability and criminal trials is a natural one. In the Middle Ages and in modern times, there existed a "mathematics" of criminal proof. Lawyers would speak of full or half proofs; they would list ways in which the items of evidence could be added or subtracted to weaken or strengthen one's case (Franklin, 2001). Some of the pioneers in the field of probability theory were interested in legal proceedings as much as they were interested in games of chance (Hacking, 1984). For instance, Jacob Bernoulli discusses the requirement for a criminal conviction in his *Ars Conjectandi*, and he writes that 'it might be determined whether 99/100 of probability suffices or whether 999/1000 is required' (Bernoulli, 1713, pt. IV). This is one of the earliest suggestions that the criminal standard of proof be equated to a threshold probability of guilt.[1] And yet, the marriage between probability and criminal trials is also a difficult one. In *On*

[1] In his *Investigations into the Probability of Judgments in Criminal and Civil Trials* (1837), Poisson was interested in understanding the behaviour of jurors, how likely they are to convict or acquit depending on the evidence being presented to them, and how the number of jurors affected the final judgment. He was also interested in calculating the probability of a defendant's guilt given the jury's judgment of guilt (Poisson, 1837).

Crimes and Punishments, for instance, Cesare Beccaria remarks that the moral certainty needed to convict is 'nothing but a probability, though a probability of such a sort to be called certainty' (Beccaria, 1764, cp. 14). This admittedly quite elusive remark seems to indicate that the criminal standard should be understood as a blend of probability and certainty.

In this chapter I address the question of whether probability alone is enough to convict. To put it more precisely, the question is whether a criminal defendant, within our conception of a just trial, can be convicted on the sole ground that his probability of guilt is sufficiently high. There have been two main responses:

> *Probabilists*: 'Certainly yes!—It is all a matter of probabilities; no evidence can establish guilt with certainty. If a sufficiently high probability of guilt is not enough to convict, will anything ever be enough?'
>
> *Traditionalists*: 'Certainly no!—It is questionable that guilt can be quantified, and even if it could, a high probability of guilt, in and of itself, cannot be enough to convict. Criminal trials are not about gambling with people's lives!'

The academic debate between probabilists and traditionalists is still ongoing today.[2]

[2] The academic debate between probabilists and traditionalists officially began in the sixties and seventies, amidst the raising popularity of the law and economics movement (Calabresi, 1961; Becker, 1968; Posner, 1973). On one hand, a number of scholars suggested that probability and statistical methods would improve our theoretical understanding of the law of evidence and procedure, e.g. one suggestion was that standards of proof be interpreted in terms of probability thresholds (Ball, 1961; Kaplan, 1968; Simon and Mahan, 1971; Lempert, 1977; Kaye, 1978). Scholars thought that probability and statistical methods would help fact-finders weigh the evidence and reach more accurate verdicts (Cullison, 1969; Finkelstein and Fairley, 1970, 1971). On the other hand, many remarked that standards of proof should not be probabilistically quantified, and that probability and statistical methods would impoverish and misconstrue legal reasoning rather than improve it (Tribe, 1971a,b; Underwood, 1977; Cohen, 1977, 1981a; Nesson, 1979).

In the eighties, the interest in questions concerning the relationship between legal evidence and probabilistic methods grew (Kaye, 1979a,b, 1980, 1982, 1986a,b,c, 1989; Saks and Kidd, 1980; Gärdenförs et al., 1983; Nesson, 1985; Allen, 1986; Lempert, 1986; Schmalbeck, 1986; Thomson, 1986; Zuckerman, 1986; Friedman, 1987; Schoeman, 1987; Thompson and Shumann, 1987; Dant, 1988; Wright, 1988; Shaviro, 1989). In 1986, a very important evidence law conference took place and its proceedings were later published in a volume under the title *Probability and Inference in the Law of Evidence* (Tillers and Green, 1988).

In the nineties, a new and powerful form of statistical evidence, DNA evidence, gained momentum

Legal scholars have been sharply divided, in particular during the sixties, seventies, and eighties. Recently, however, Peter Tillers (2011) has noted that the disagreement between probabilists and their opponents need not be as dramatic as it seems. I agree. So, instead of taking sides, my plan is to clarify, distinguish, refine, and reconcile. I offer a refinement of legal probabilism, which I call *new probabilism*, and I contend that new probabilism can withstand many criticisms from the traditionalist side.

The plan is as follows. In sections 2.1 and 2.2, I begin with a statement of legal probabilism and I lay out some of the arguments for and against. In section 2.3, I offer a first refinement of legal probabilism, which I call *qualified probabilism*. The key idea is the distinction between error reduction and error distribution. In section 2.4, I argue that the criminal standard of proof serves both the function of error distribution and of error reduction. In section 2.5, I offer my final refinement of legal probabilism, *new probabilism*. In the final section, I argue that new probabilism can withstand many traditionalist criticisms.

2.1 LEGAL PROBABILISM

To begin with, legal probabilism can be seen as the conjunction of two claims:

> QUANTIFICATION CLAIM: a probabilistic quantification of the defendant's guilt can be given through an appropriate weighing of all the fallible evidence available (that is, of all the evidence against, and of all the evidence in defense of, the accused);
>
> THRESHOLD CLAIM: an appropriately high threshold guilt probability, say, 0.9 or 0.99, should be the decision criterion for criminal convictions.

in the courtroom. The nineties were characterized by less ideological, and more empirical, analyses of the role of statistics and probability in the courtroom (Koehler and Shaviro, 1990; Allen, 1991; Fienberg and Straf, 1991; Feinberg and Kaye, 1991; Robertson and Vignaux, 1993; Schum and Kadane, 1996; Dekay, 1996; Kaye, 1999; Taroni and Aitken, 1997). Unsurprisingly, many articles were published on DNA evidence, its legal and statistical pros and cons (Kaye, 1993; Koehler, 1993; Koehler et al., 1995; Koehler, 1996a; Lempert, 1993; Robertson and Vignaux, 1995; Balding and Donnely, 1996). Finally, in the 21st century the interest in statistical and probabilistic methods in the courtroom grew significantly. In particular, two recent monographs on the foundations of evidence law devoted entire chapters on the topic (Stein, 2005; Ho, 2008).

In this chapter, I will be mostly concerned with the threshold claim, but it is nevertheless important to appreciate the significance of the quantification claim. We can interpret the quantification claim in at least two different ways. We can interpret it as the claim that a quantification of guilt—understood as an actual reasoning process—can be effectively carried out by the fact-finders. We can also think of the quantification claim as expressing a mere idealization or a regulative ideal. The latter interpretation seems more plausible. It has become clear, after all, that it is unrealistic to effectively quantify the probability of guilt, and the probabilists themselves have come to admit that. For instance, the authors of a recent book on probabilistic inference in forensic science write that 'the ... [probabilistic] formalism should primarily be considered as an *aid to structure and guide one's inferences under uncertainty, rather than a way to reach precise numerical assessments*' (Taroni et al., 2006, p. xv). Even from a probabilist standpoint, then, the quantification of guilt is an idealization which has, primarily, a heuristic role.

Just as the quantification claim can be interpreted in two different ways, the same can be said of the threshold claim. For one, we can interpret it as describing an effective decision procedure, as though the fact-finders were required to mechanically convict whenever the defendant's probability of guilt happened to meet the desired probabilistic threshold. But there is a second, and less mechanistic, interpretation of the threshold claim. On the second interpretation, the threshold claim would only describe a way to understand, or theorize about, the criminal standard of proof or the rule of decision in criminal trials. (Incidentally, as a matter of terminology, I will use the expressions 'criminal standard of proof,' 'rule of decision in criminal trials,' and 'decision criterion in criminal trials' interchangeably; they all refer to the set of conditions that are sufficient for a criminal conviction.) The second interpretation of the threshold claim—which fits well with the "idealization interpretation" of the quantification claim—is the one which many legal probabilists endorse, and it is less likely to cause outrage among the traditionalists.[3] It is also the interpretation I shall adopt for the purpose of this chapter.

[3] Lawrence Tribe, in his famous 1971 article on trial-by-mathematics, expresses disdain for a trial process that were to be mechanically governed by numbers and probabilities. He claims that in this case the jurors would forget their humanizing function. He writes:

> Guided and perhaps *intimidated by the seeming inexorability of numbers*, induced by the

Different interpretations aside, I will now focus on an argument in defense of the threshold claim. It goes as follows:

> (a) The occurrence of some decisional errors in criminal trials is *inevitable*.
>
> (b) There are *two types* of errors, i.e. convicting an innocent and acquitting a guilty defendant.
>
> (c) Given (a) and (b), the rule of decision in criminal trials can only (or primarily) have the function of *distributing* errors in a desirable way.
>
> (d) A desirable error distribution can be promoted, albeit not guaranteed, by setting an appropriate *threshold* guilt probability for criminal convictions. A higher threshold lowers the rate of wrongful convictions, while a lower threshold increases the rate of wrongful convictions.
>
> (e) Given (d), the threshold guilt probability for criminal convictions should be set to an appropriately high value because this promotes a desirable error distribution, i.e. a distribution with proportionally fewer wrongful convictions. (The assumption here is that convicting an innocent is more harmful than acquitting a guilty.)
>
> (Conclusion) The threshold claim holds because—as (c) asserts—distributing errors in a desirable way is the criminal standard's only (or primary) function, and—as (e) asserts—an appropriately high threshold performs precisely that function.

Call this argument the *probabilist threshold argument*. Let's now look at each step more closely. The first three steps are relatively straightforward. Indeed, (a) some decisional errors are inevitable, because the trial system is not infallible, and (b) there are two kinds of error, i.e. wrongful convictions and wrongful acquittals.[4] The inevitability of

> persuasive force of formulas and the precision of decimal points to perceive themselves as performing a largely mechanical and automatic role, *few jurors ... could be relied upon to recall, let alone to perform, [their] humanizing function*. (Tribe, 1971b, emphasis mine)

But one can see that this worry does not apply if we interpret the threshold claim in a non-mechanistic way.

[4] Quite aptly, Justice Harlan in his concurring opinion in *re Winship* (1970) writes:

> [...] the trier of fact will sometimes, despite his best efforts, be wrong in his factual conclusions ... [A] factual error can make a difference in one of two ways. First, it can result in ... the conviction of an innocent man. On the other hand, an erroneous factual

errors, of one kind or another, puts an upper limit on the trial system's ability to reduce errors, so that the system can only aim at a desirable error distribution. The distinction between error reduction and error distribution is crucial here. Error reduction means to reduce the overall rate of wrongful acquittals *and* wrongful convictions, while error distribution means to instate or promote a certain *proportion* of wrongful convictions to wrongful acquittals.[5] Once errors cannot be further reduced, the rule of decision in criminal trials can only have the function of distributing errors according to some criterion of social desirability or acceptability. This is what step (c) asserts. (In section 2.3 the distinction between error reduction and error distribution is spelled out more precisely, and a stronger argument for step (c) is presented.)

The next step in the argument—step (d)—requires a more elaborate explanation. Suppose that, based on the evidence available, a probability of guilt can be associated with each defendant: the stronger the incriminating evidence (and the weaker the exculpatory evidence), the higher the probability of guilt. The key point here is that the long run distribution of errors can be influenced by moving the probabilistic threshold for a conviction upwards or downwards. Other things being equal, a more demanding threshold will promote—in the long run—proportionally fewer wrongful convictions (and proportionally more wrongful acquittals); a less demanding threshold will do the opposite. To see why, suppose a higher probabilistic threshold for criminal conviction is enforced. Fewer defendants will then be convicted in the long run, because fewer defendants will meet the higher threshold compared to a lower threshold. And among those defendants who escape conviction as a result of the higher threshold, there will be both innocent and guilty defendants. How so? We should bear in mind that, given

determination can result in . . . the acquittal of a guilty man. 397 U.S. 358, 370.

[5] The distinction between error reduction and error distribution is taken from (Laudan, 2006). Instead of error distribution, some authors speak of error allocation or allocation of the risk of error; see e.g. (Stein, 2005). Error allocation is a slightly different concept from error distribution. Error allocation has to do with which party at trial—defense of prosecution—would bear the burden if a certain error were to occur. Although one might argue that wrongful convictions burden defendants and wrongful acquittals burden prosecutors, the concept of error allocation brings in a further level of complexity which I do not discuss here.

the fallibility of the system, the probability of guilt of some innocent defendants will be as high as that of some guilty defendants, and conversely, the probability of guilt of some guilty defendants will be as low as that of some innocent defendants. So, a higher threshold will have two concomitant effects. For one, fewer *innocent* defendants will be convicted, and for another, fewer *guilty* defendants will be convicted (which means that more guilty defendants will be acquitted).[6] All in all, a higher threshold will have the effect of reducing the rate of wrongful convictions (at the cost of increasing the rate of wrongful acquittals); a lower threshold, on the other hand, will have the opposite effect. This establishes step (d) in the argument—i.e. that a certain error distribution can be promoted by adopting a certain probabilistic threshold. (For another argument in support of step (d), see section 2.3.)

I should emphasize that setting a certain threshold for a conviction does not automatically guarantee a certain distribution of wrongful convictions and wrongful acquittals. For it is not the case that if the threshold is, say, 0.9, then the resulting distribution (or

[6] Here is an example. Let us consider a very simple scenario involving nine criminal trials and nine defendants. Some of them are factually guilt, say six, and some of them are innocent, say three. Each defendant is associated with a probability of guilt on the evidence, as represented by the following table:

DEFENDANT	GUILT PROBABILITY	FACTUALLY GUILT?
1	0.9	Yes
2	0.8	Yes
3	0.8	Yes
4	0.7	Yes
5	0.6	No
6	0.6	Yes
7	0.5	Yes
8	0.4	No
9	0.3	No

Note that some factually innocent defendants are associated with the same probability of guilt as some guilty defendants (or even with a higher probability). This reflects the fact that some errors are inevitable. (If we could tell without error who is guilty and who is innocent just by looking at each defendant's probability of guilt, the trial system would be infallible. But, clearly, it is not.) Now suppose that the threshold is set to a relatively low value, say 0.6. The decisional errors in our simple scenario would then amount to one wrongful conviction and one wrongful acquittal. Suppose, instead, that the probabilistic threshold is set to a higher value, say, 0.8. The decisional errors would now amount to zero wrongful convictions and three wrongful acquittals. So, a higher threshold has the effect of reducing the number of wrongful convictions (in our example, from one to zero) at the costs of increasing the number of wrongful acquittals (in our example, from one to three); a lower threshold has the opposite effect.

relative proportion, or error ratio) will be 1:10. There could be other factors besides a threshold for criminal convictions which may influence the distribution of errors, e.g. the base rate proportion of innocent to guilty defendants among those who face trial,[7] or the difference between the expected and the actual distribution of errors.[8] Consequently, it is more appropriate to say—as my formulation of step (d) does—that setting a threshold is a way to promote, rather than guarantee, a certain error distribution.

Next, step (e) in the argument rests on a widely accepted idea: a wrongful conviction is more harmful than a wrongful acquittal. This means that, in the long run, it is more desirable to have fewer wrongful convictions (at the cost of having more wrongful acquittals) than having fewer wrongful acquittals (at the cost of having more wrongful convictions). This preference for proportionally fewer wrongful convictions, together with the fact that the rate of wrongful convictions can be reduced by setting a higher threshold, yields the conclusion that (e) the threshold guilt probability *should* be set to an appropriately high value.[9]

The argument is not complete yet. Step (e) alone is insufficient to establish the threshold claim. Step (e) asserts that the threshold for criminal convictions should be set to an appropriately high value. The threshold claim, instead, asserts that an appropriately high threshold probability should be the rule of decision or the standard of proof in criminal trials. There is a missing link between (e) and the threshold claim. The missing

[7] The importance of base rate proportions can be illustrated with probabilities. Let I and G be short for 'the defendant is innocent' and 'the defendant is guilty.' Let A and C be short for 'the defendant is acquitted' and the 'the defendant is convicted.' Let $C \wedge I$ and $A \wedge G$ be short for 'an innocent defendant is convicted' and 'a guilty defendant is acquitted.' In probabilistic terms, the (expected) distribution of errors can be expressed by the ratio $\frac{P(C \wedge I)}{P(A \wedge G)}$. The probability calculus tells us that $\frac{P(C \wedge I)}{P(A \wedge G)} = \frac{P(C|I)}{P(A|G)} \times \frac{P(I)}{P(G)}$. From the equality it is clear that the (expected) distribution of errors, i.e. $\frac{P(C \wedge I)}{P(A \wedge G)}$, depends on the (expected) base rate distribution of innocent and guilty defendants, i.e. $\frac{P(I)}{P(G)}$. For a nuanced discussion on these and similar issues, see (Dekay, 1996).

[8] Since we are working with probabilities, we are not working with actual error distributions, but only with expected, long run error distributions (Kaye, 1999).

[9] What is the justification for using normative language here? A justification can be given through the following rationality principle: if an end is desirable, then the means towards that end *should* be intended (unless there are obstacles or higher priorities at stake). An appropriately high threshold guilt probability is a means toward the desirable end of having proportionally fewer wrongful convictions than wrongful acquittals. So, the threshold guilt probability *should* be appropriately high.

link, however, can be filled by appealing to the earlier step (c) in the threshold argument. Step (c) asserts that the only (or primary) function of the criminal standard of proof is to promote a desirable error distribution. It is the standard's only function because some errors are inevitable, and thus what the standard can do—at best—is to distribute errors appropriately. We can read step (c) as the statement that distributing errors in a desirable way characterizes the "essence" of the criminal standard of proof. If so, whichever evidentiary or procedural constraint performs the function of distributing errors in a desirable way must serve as the criminal standard. An appropriately high threshold probability, then, can serve as the criminal standard of proof because it has precisely the function of promoting a desirable error distribution.[10] This establishes the threshold claim.[11]

[10] A desirable error distribution could be promoted by means other than an appropriately high threshold probability of guilt; call such other means Y. Whether such Y exists or not should not concern us here. But if such Y exists, Y would serve as the rule of decision in criminal trials according to step (c). This does not contradict the conclusion that an appropriately high threshold probability can *also* serve as the rule of decision. What serves as the rule of decision, then, need not be unique.

[11] The threshold claim can also be derived from statistical decision theory according to which the trial system should convict whenever convicting maximizes expected utility. We have seen that two errors affect criminal trials, i.e. convicting an innocent and acquitting a guilty defendant (abbreviated CI and AG). Suppose we can assign a probability to guilt and innocence (abbreviated $P(G)$ and $1 - P(G)$). Further, suppose we can assign a (negative) utility u to CI and AG. (Utilities can have negative values.) On a simplified model, a conviction maximizes expected utility provided:

$$1 - P(G) \times u(CI) > P(G) \times u(AG).$$

The inequality expresses the fact that the expected (negative) utility resulting from convicting an innocent is higher than the expected (negative) utility resulting from acquitting a guilty defendant. So, even in light of possible errors, convicting turns out to be better as far as expected utility goes. The inequality holds just in case:

$$\frac{1 - P(G)}{P(G)} > \frac{u(AG)}{u(CI)}.$$

Suppose $\frac{u(AG)}{u(CI)} = \frac{-1}{-99}$, i.e. convicting an innocent defendant is deemed far worse than acquitting a guilty defendant. It follows that $\frac{1-P(G)}{P(G)} > \frac{1}{99}$, so the probability of guilt must be greater than 0.99. As the ratio $\frac{u(AG)}{u(CI)}$ varies, the threshold guilt probability will have to change accordingly. (This argument can also be extended to include more variables, e.g. the (positive) utilities associated with convicting a guilty defendant and acquitting an innocent.) Importantly, statistical decision theory can establish the threshold claim insofar as it assumes that the only function of the criminal standard is that of promoting the system's compliance with the ratio $\frac{u(AG)}{u(CI)}$. This assumption parallels step (c) in the threshold argument. Promoting

The threshold argument in support of the threshold claim is by no means uncontroversial. But all in all, it makes an important point. Given that there is an inherent component of fallibility and error in legal-decision making, how can the system cope with it? The probabilists think that all the system can do is to promote a desirable distribution of errors, and that setting an appropriate threshold guilt probability as the rule of decision is an effective way to do so. How will the traditionalists respond?

2.2 BACK AND FOURTH

If you recall, in my statement of legal probabilism, I distinguished the quantification claim and the threshold claim. The former asserts that a probabilistic quantification of guilt is possible; the latter asserts that an appropriate probabilistic threshold for criminal convictions should be the rule of decision in criminal trials. In critiquing legal probabilism, the traditionalists can begin by challenging the quantification claim. They can question whether we can quantify guilt at all.[12] Studies in psychology suggest that, given the evidence presented during a trial, fact-finders construct stories and coherent narratives of what happened (Pennington and Hastie, 1991; Simon, 2004; Devine, 2012). Fact-finders, according to these studies, do not engage in an abstract process of probability assessment. But if a probabilistic quantification of guilt or innocence is far removed from what goes on in the minds of the fact-finders, the traditionalists may well wonder whether it makes sense to even begin an academic conversation about it.

Although quantifying guilt is a psychologically implausible process, it can still make sense as an idealization that guides the fact-finders' assessment and weighing of the

compliance with the ratio $\frac{u(AG)}{u(CI)}$, after all, is very much like promoting a desirable error distribution.

[12] For one thing, the traditionalists must acknowledge that it is not at all implausible to quantify the probative value of those pieces of evidence which have an overtly statistical or numerical form. This can be done, in a circumscribed manner, by saying that a certain statistics-based item of evidence supports a proposition—e.g. the defendant left his genetic traces at the crime scene—with a certain probability. Yet, the complaint of the traditionalists still stands, because quantifying the probative value of statistics-based evidence is not the same as quantifying the probative value of evidence that is overtly non-quantitative. And the further difficulty is that quantifying the probative value of some items of evidence relative to a circumscribed proposition is different from quantifying guilt or innocence as such.

evidence. Quantification, in other words, can have a regulative and heuristic role; and this is how I suggest we interpret the quantification claim. Some traditionalist will, of course, still disagree. They will argue that legal reasoning is non-probabilistic and that any attempt to bring probability into the picture—even as a heuristic device—is doomed to obscure our understanding of criminal trials. These are important challenges, but I cannot address them here. My focus will be on the threshold claim rather than the quantification claim. As we shall soon see, even if we confine ourselves to the threshold claim, there are plenty of controversial issues that need addressing. The disagreement between probabilists and traditionalist, after all, runs deeper than the question of whether or not guilt can be quantified.

So let's grant the probabilists that, with some robust dose of idealization, a quantification of guilt is possible. Turning now to the threshold claim, the probabilists think that the rule of decision in criminal trial amounts to a threshold guilt probability, and they offer the threshold argument in support of this claim. How good of an argument is it? Here is a powerful objection from the traditionalist side:

> RANDOM ACQUITTAL OBJECTION. The probabilists argue—in step (e) of their threshold argument—that the threshold for a conviction should be set to an appropriately high value because this promotes a desirable distribution of errors. But here is an obviously unacceptable policy which also promotes a desirable error distribution. Suppose we adopt the policy of randomly letting some defendants go free every month. On the plausible assumption that those randomly acquitted include some guilty defendants, the policy would promote an absolute increase in the rate of wrongful acquittals, which, in turn, would promote proportionally fewer wrongful convictions relative to wrongful acquittals. Yet, many would be uneasy—even the probabilists—in saying that we *should* pursue this policy because it promotes the desirable goal of having proportionally fewer wrongful convictions. And if so, it becomes not at all clear that we should set the probabilistic threshold for criminal convictions to a high value.[13]

The objection shows a couple of things: first, if a policy promotes a desirable error distribution, this alone is not a good justification for adopting the policy; second, it is

[13] Laudan (2008) offers an argument along the lines of the random acquittal objection.

discomforting to promote a desirable error distribution if this requires a gratuitous or deliberate increase in the overall rate of error. The random acquittal objection undermines the motivation for setting the threshold probability to a high value. If the motivation is a desirable error distribution, the objection shows that there are ways to promote a desirable error distribution which are intuitively unacceptable, such as randomly acquitting some defendants. This undermines step (e) in the threshold argument.

The legal probabilists must acknowledge that there is something wrong with the threshold argument. And since the threshold claim rests on the threshold argument, the random acquittal objection is a serious challenge to the threshold claim as well. But even granting that the threshold argument contains some weaknesses, the probabilists can defendant the threshold claim more directly, as follows: 'If you traditionalists disagree that a high threshold probability of guilt is enough to convict, what else would be enough?' In the traditionalist camp, Lawrence Tribe (1971b) attempted to answer this difficult question. I concentrate on his view because it neatly highlights what is at stake. Tribe thought that the trial system should strive for certainty, not absolute certainty, of course, but '*as close an approximation to certainty as seems humanly attainable*' (Tribe, 1971b, p. 1374). So we have:

> TRIBE'S CERTAINTY. The criminal justice system should strive for certainty. This need not be absolute, but it should be as close an approximation to certainty as humanly possible. This is what the criminal standard of proof should be: *humanly attainable certainty of guilt*.

Such a demand for certainty is not Tribe's own creation. The expression 'moral certainty'—sometimes used to explicate the meaning of the criminal standard of proof—hints at the same demand for a humanly attainable certainty. Recall, also, Beccaria's remark that the moral certainty needed to convict is 'nothing but a probability, though a probability of such a sort to be called certainty' (Beccaria, 1764, cp. 14). Tribe's demand for certainty, however, is unlikely to move the probabilists, who can respond as follows:

> THE STOPPING OBJECTION. Tribe says that the criminal justice system should strive for certainty, but he must admit that this "striving" stops at some point; *it cannot go on forever*. But when can it stop? Which criterion

> decides when to stop? The criterion can be full certainty or a high threshold probability of guilt. The first is obviously unworkable, so the second criterion is the only viable option.

The challenge for the traditionalists is to offer a criterion which determines when the demand for certainty is effectively met. If we follow the logic of the stopping objection, the criterion in question can be *either* full certainty *or* a high threshold guilt probability. The force of the stopping objection is that there does not seem to be any other plausible criterion besides full certainty (which is unattainable) or a threshold probability (which is the probabilist criterion). The traditionalists seem to be stymied on this issue.

I should emphasize that the stopping objection is stronger than it might appear at first. On behalf of the traditionalists, the trained epistemologist might be tempted to dismiss the objection outright on the conviction that it is based on the false dichotomy between certainty of guilt and high probability of guilt. Epistemologists within the analytic tradition, in fact, have offered several accounts of knowledge which neither reduce it to mere high probability nor equate it to certainty.[14] Tribe's demand for certainty, the trained epistemologist might continue, can be understood as a demand for knowledge (according to one's own preferred account of knowledge).[15] I want to explain why this route is not straightforward by considering one account of knowledge on offer. A popular view states that the knowledge of a proposition p requires that one's evidence be capable of ruling out all alternative scenarios which make p false. Importantly, the alternatives to be ruled are not *all* the alternatives: if that were the case, knowledge would again amount to certainty. The alternatives to be ruled out must belong to a suitable

[14] Some argue that knowledge requires that our evidence should be causally or explanatorily connected with the known proposition; see Nelkin (2000). Other invoke modal notions such as "safety" and "tracking," and argue that in order to count as knowledge a belief should be safe or tracking; see Pritchard (2005), Williamson (2000), DeRose (1996), and Roush (2006). Still others argue that in order to know we should be a position to rule out certain relevant alternative scenarios; see Dretske (1971), Vogel (1999) and Lewis (1996). Note that these accounts of knowledge do not require certainty as a condition for having knowledge of a proposition.

[15] The connection between knowledge and criminal convictions was first explored by Thomson (1986). Enoch et al. (2012) defend a "sensitivity view" of knowledge and apply it to trial proceedings. Redmayne (2008) offers an excellent survey of the relationship between probability, trial proceedings, and knowledge.

set, sometimes defined as the set of relevant or reasonable alternatives (Dretske, 1971; Lewis, 1996; Lawlor, 2013). This “rule-out-alternatives” view of knowledge echos the criminal standard of proof. The latter requires that guilt be proven beyond a reasonable doubt, namely that all reasonable doubts, or all reasonable alternatives to guilt, be ruled out. But this is where a daunting difficulty arises. How are we to understand a reasonable doubt, as opposed to an unreasonable one? The stopping objection, then, issues a non-trivial challenge for the trained epistemologist. If we were to develop the parallelism between knowledge and warranted conviction, the challenge would be to offer an account that does not reduce knowledge to mere high probability nor certainty, an account, *in addition*, which can explain the notion of reasonable doubt and relevant alternative in a manner that is more informative than the formula ‘guilt beyond a reasonable doubt.’ This is by no means an easy task.

I will not survey the existing accounts of knowledge and see which can best answer the stopping objection. This would lead us too far from criminal trials. I rather prefer to follow more closely the dialectic between probabilists and traditionalists, and then attempt to offer a reconciliation of the two sides. To this end, let us take a step back and recapitulate what we have so far. I offered a reconstruction of the disagreement about the threshold claim, i.e. the claim that an appropriately high probability of guilt is enough to convict. My reconstruction went as follows. The probabilists first formulated the threshold argument in support of the threshold claim. According to the threshold argument, the criminal standard of proof can only control the distribution of errors, and since a wrongful conviction is more harmful than a wrongful acquittal, the standard should be equated to an appropriately high threshold guilt probability. The traditionalists then responded with the random acquittal objection: if the motivation to set the threshold to a high value is that it promotes a desirable error distribution, the policy of randomly acquitting some defendants has also the effect of promoting a desirable error distribution, yet this does not mean we should endorse such a policy. This objection puts pressure on the threshold argument and the threshold claim. Next, when the traditionalists were asked to give their own decision criterion, they invoked something like Tribe’s humanly attainable certainty. In response, the probabilists challenged Tribe’s certainty by means of the stopping objection: humanely attainable certainty as the rule of decision in criminal trials cannot be any different from an appropriately high threshold guilt probability.

So we have four, all seemingly persuasive, pieces of reasoning: the probabilist threshold argument (from the probabilist side); the random acquittal objection (from the traditionalist side); Tribe's certainty (again from the traditionalist side); the stopping objection (from the probabilist side). They pull us in different directions, either towards the threshold claim or away from it, and we cannot concurrently maintain all of them. What are we to do? My suggestion is that, in order to gain some clarity here, we should appreciate the distinction between error reduction and error distribution, and how it informs the disagreement between traditionalists and probabilists. In the next section, I will argue that the traditionalists are too narrowly focused on error distribution, and that they have overlooked the importance of error reduction. This will afford us a way to address both the stopping objection and the random acquittal objection, while at the same time holding on to the threshold argument and Tribe's certainty.

2.3 Error distribution and error reduction

The probabilists interpret the criminal standard as a threshold probability because they understand the standard's function to be that of promoting a desirable distribution of decisional errors. The preference for having proportionally fewer wrongful convictions mandates that the threshold be set to a high value. Importantly, a high threshold guilt probability does *not* have the function of reducing decisional errors, and in fact, it has the opposite effect: it increases the overall rate of errors. To illustrate this claim, I shall make use of Signal Detection Theory (SDT), a theoretical framework which can model how well a system performs at detecting signals, such as the presence of an obstacle, a prey, an imminent danger, a desease, etc. Many authors have already noted how SDT can be applied to the trial context (Underwood, 1977; Dekay, 1996; Wickens, 2002). Here is how the analogy goes. The signal to be detected would be factual guilt or factual innocence. The system by which guilt and innocence are detected would be the trial system itself, and more precisely, the strength and weakness of the evidence, which can be measured in guilt probability. The SDT framework is very useful for appreciating the distinction between error distribution and error reduction, so I shall use it throughout this section.

Consider now the diagram below in figure 2.1 which represents—over a large number of criminal trials—how guilty and innocent defendants are distributed around certain guilt probability values.

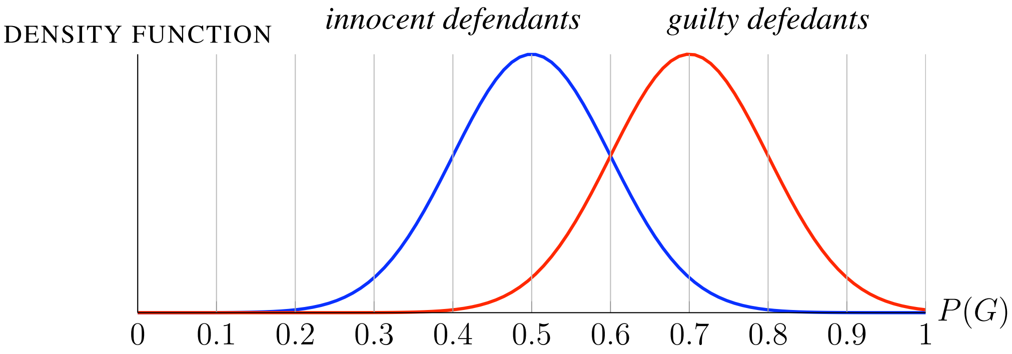


Figure 2.1: The *x*-axis represents the guilt probability on the evidence (sometimes called the "apparent guilt"). The *y*-axis represents the percentage of defendants (factually innocent or guilty) within a certain range of guilt probabilities.

There are two overlapping curves in the diagram. The left curve represents the distribution of factually innocent defendants relative to guilt probability values; the right curve does the same for guilty defendants. Note that the areas below the two curves partly overalp, so that the guilt probability of some innocent defendants can appear as high as the guilt probability of some innocent defendants. The trial system represented in the diagram, in other words, is fallible, as we would expect it to be. But from the diagram we can also see that most innocent defendants have a guilt probability around 0.5, while most guilty defendants have a guilt probability around 0.7. So the diagram conveys an important difference between innocent and guilty defendants: guilty defendants have, on average, higher guilt probabilities than innocent defendants. This is what we would expect from a well-functioning, albeit still fallible, trial system.[16]

So far I have not introduced any decision criterion for acquittals and convictions. Let us then stipulate that the decision criterion equals a 0.7 probabilistic threshold. In other words, the criterion is that if a defendant's guilt probability meets the 0.7 threshold, the system convicts the defendant, otherwise it acquits, as represented in the diagram in figure 2.2. The two shaded areas, which are divided by the 0.7 threshold line, represent the percentage of defendants who are wrongfully acquitted and the percentage of defendants who are wrongfully convicted. These percentages can also be expressed in terms of conditional probabilities, i.e. the probability that a defendant is acquitted given his guilt, $P(A|G)$, and the probability that a defendant is convicted given his innocence, $P(C|I)$.

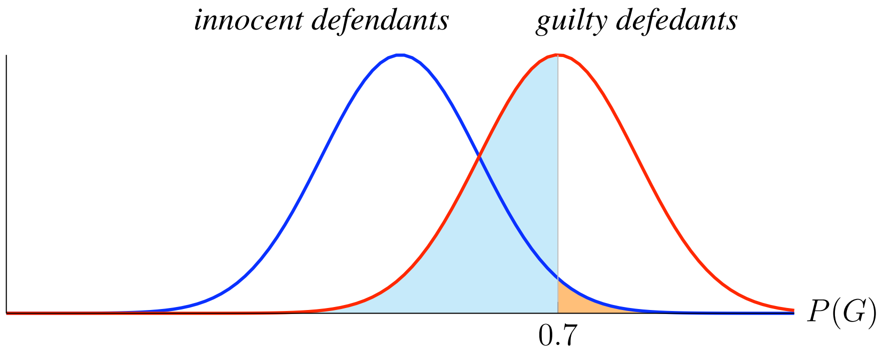


Figure 2.2: The vertical line at 0.7 represents the threshold line. The defendants who have a probability of guilt higher than 0.7 are convicted, otherwise they are acquitted.

Now compare the above diagram with the diagram in figure 2.3. The threshold is now set to a slightly lower value, 0.65, while everything else remains the same. There

[16] Three clarifications. First, for convenience, both distributions are normal. This means that, in the long run, few defendants have extremely high or extremely low guilt probabilities, and most of them are concentrated around the average. Second, the total area underneath each curve is 100 percent (of guilty defendants) and 100 percent (of innocent defendants), regardless of how many innocent or guilty defendants there are. This means that the relative size of innocent and guilty defendants cannot be inferred from the areas underneath the two curves. Third, the two curves have the same spread. This need not be the case. It may well be that the guilt probabilities of innocent defendants are more spread out than the guilt probabilities of innocent defendants, or vice versa.

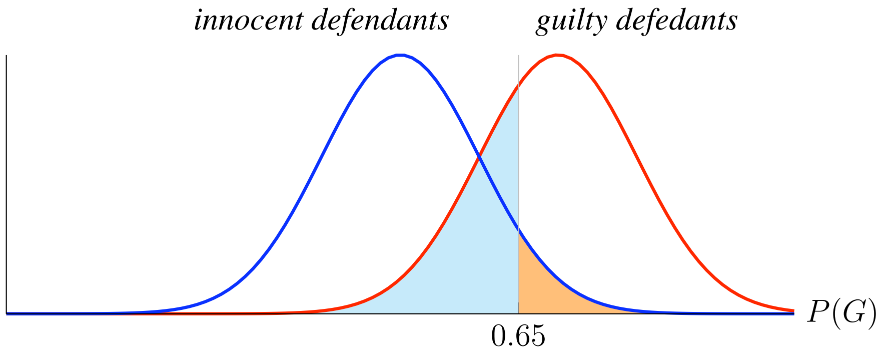


Figure 2.3: The vertical line at 0.65 represents the threshold line. The defendants who have a probability of guilt higher than 0.65 are convicted, otherwise they are acquitted.

are two notable differences between the two diagrams. One difference concerns the distribution of errors. The rate of wrongful convictions is higher in the diagram with the lower threshold. Visually, the shaded area representing the percentage of wrongful convictions—the shaded area to the right of the threshold line—is bigger in the diagram with the lower threshold. At the same time, the rate of wrongful acquittals is lower in the diagram with the lower threshold. Visually, the shaded area representing the percentage of wrongful acquittals—the shaded area to the left of the threshold line—is smaller in the diagram with the lower threshold. So, comparing the two diagrams makes it clear that, other things being equal, the more demanding the threshold, the fewer wrongful convictions (and the more wrongful acquittals). This is, once again, step (d) of the threshold argument from section 2.1.

The second notable difference between the two diagrams concerns the overall error rate. In the diagram with the lower threshold, the total shaded area representing both wrongful convictions and wrongful acquittals has shrunk, although the shaded area on the right—the one corresponding to the rate of wrongful convictions only—has widened. This means that by lowering the threshold (up to a certain point at least), the total shaded area representing both wrongful acquittals and wrongful convictions will shrink. *So the overall rate of errors decreases by lowering the threshold probability for criminal* *convictions (up to a certain point at least)*.[17] This conclusion shows quite clearly that an appropriately high threshold guilt probability does *not* have the function of reducing errors; it only has the function of distributing errors in a desirable way.

***

What I have said so far should not surprise the probabilists. They are well aware that their theory assigns to the criminal standard the function of distributing errors, not the function of reducing errors. This is very clearly stated in step (c) of the threshold argument. But in light of the contrast between error distribution and error reduction, it is tempting to rephrase step (c) as follows:

> (c*) The rule of decision in criminal trials can only (or primarily) have the function of *distributing* errors in a desirable way *provided errors have been reduced as much as possible.*

This would still be a way to hold on to step (c). If it is indeed the case that errors have been reduced as much as possible, it follows that all is left to do is to distribute them appropriately, in agreement with the probabilist point. The formulation in (c*) is attractive for another reason. It can be used to offer a workable interpretation of Tribe's demand that criminal trials strive for the greatest possible certainty. The unworkable interpretation of Tribe's demand for certainty is that guilt should be proven with a one hundred percent probability; this is unworkable because the certainty of guilt is hardly, if ever, attainable. But if we take inspiration from (c*), Tribe's certainty can be interpreted differently, namely as the demand that errors be reduced as much as possible. If so, (c*) seems to represent a middle ground between the probabilist position and the traditionalist position.

In order for (c*) to be well formulated, however, the notion of error distribution and error reduction have to be properly distinguished. Unfortunately, this is not entirely possible, and thus (c*) suffers from a serious flaw. As seen earlier, distributing errors in

[17] If the threshold becomes too low, however, the overall error rate will increase again. The argument that a lower threshold contributes to error reduction is spelled out more generally in (Kaye, 1982).

a desirable way requires an increase in the overall rate of error. A demanding threshold guilt probability can promote a desirable rate of errors, yet this also causes the overall error rate to rise (at least in comparison to a lower threshold). So, it makes little sense to say that we should reduce errors as much as possible *and* that we should distribute them appropriately. Reducing errors as much as possible and distributing them in a desirable way are inconsistent goals, and the formulation of (c*) is at best incoherent. We have:

> INCOMPATIBILITY CLAIM. Reducing errors as much as possible conflicts with distributing them in a desirable way (i.e. having proportionally fewer wrongful convictions relative to wrongful acquittals).

The incompatibility claim—as the probabilists will be happy to hear—gives us one more reason to endorse step (c) and discard (c*). If the goals of error reduction and error distribution are inconsistent, the criminal standard of proof can be in charge of *either* error distribution *or* error reduction, not both. The probabilists have chosen error distribution. This choice is in agreement with the widespread intuition that the point of the criminal standard is to make convicting hard, and this has mostly to do with error distribution rather than error reduction.

Despite its manifest incoherence, (c*) expresses an intuitive idea: we would like a trial system which reduces errors as much as possible *and* distributes them appropriately. Unfortunately, the incompatibility claim seems to make this impossible. But is there a way to amend (c*) and retain its intuitiveness? We need a more nuanced way to talk about error reduction so that no manifest incoherence arises. This is what I set out to do next by elaborating further on the notion of error reduction and showing that we can work around the incompatibility claim.

To begin with, consider again the diagrams on page 54–55. If you recall, the shaded area to the right of the threshold line represents the percentage of innocent defendants who are convicted, and the shaded area to the left of the threshold line represents the percentage of guilty defendants who are acquitted. As noted earlier, these percentages can be expressed in terms of conditional probabilities:

> $P(C|I)$, i.e. the probability that one is convicted given that one is innocent;

$P(A|G)$, i.e. the probability that one is acquitted given that one is guilty.

Aiming to reduce the overall rate of error in criminal trials can mean two things: (i) reducing *both* $P(C|I)$ and $P(A|G)$; (ii) reducing one probability at the costs of increasing the other provided the reduction of one probability offsets the increase of the other.[18] We have already seen a way to achieved error reduction in terms of (ii). If the threshold guilt probability is lowered, say, from 0.7 to 0.65, the result is a lower rate of wrongful acquittals at the costs of a higher rate of wrongful convictions. The decrease in wrongful acquittals, however, is more pronounced than the increase in wrongful convictions, so the former offsets the latter.[19] The overall error rate thus decreases.

Is there a way to reduce the error rate which takes the form of (i)? Clearly, if there is such a way, it cannot consist in moving upwards or downwards the threshold guilt probability, because if the threshold is rendered more demanding, the result is a lower rate of wrongful convictions at the cost of a higher rate of wrongful acquittals, and if the threshold is rendered more lenient, we have the opposite effect. No change in threshold can lead to a reduction of *both* error rates. Compare now the two diagrams in figure 2.4 on the next page. Although the threshold guilt probability has been kept to 0.7, the shaded areas in the second diagram have shrunk, so the overall error rate has decreased. What has changed in the second diagram is that the distance between the averages of the two curves has increased. Visually, the curves representing innocent and guilty defendants have been pushed further away from each other, so that the area of overalp has diminished. Within the SDT framework, the distance between (the averages of) the curves is sometimes called the *accuracy* of the system.[20] I prefer to speak of *discriminating power* of the trial system because the term 'accuracy' invites confusion and misunderstanding. My terminological choice should be intuitive enough. The moving away of

[18] There is a third way, i.e. to reduce the number of innocent defendants and increase the number of guilty defendants. I do not consider this option because selecting defendants is more the business of pretrial proceedings rather than trial proceedings.

[19] To see how the decrease in wrongful acquittals can offset the increase in wrongful convictions, compare the two diagrams on page 54–55.

[20] Let $N(\mu_i, 1)$ and $N(\mu_g, 1)$ denote the normal distributions for innocent and and guilty defendants respectively, with μ_i and μ_g being their averages. The system's accuracy is the difference between μ_i and μ_g.

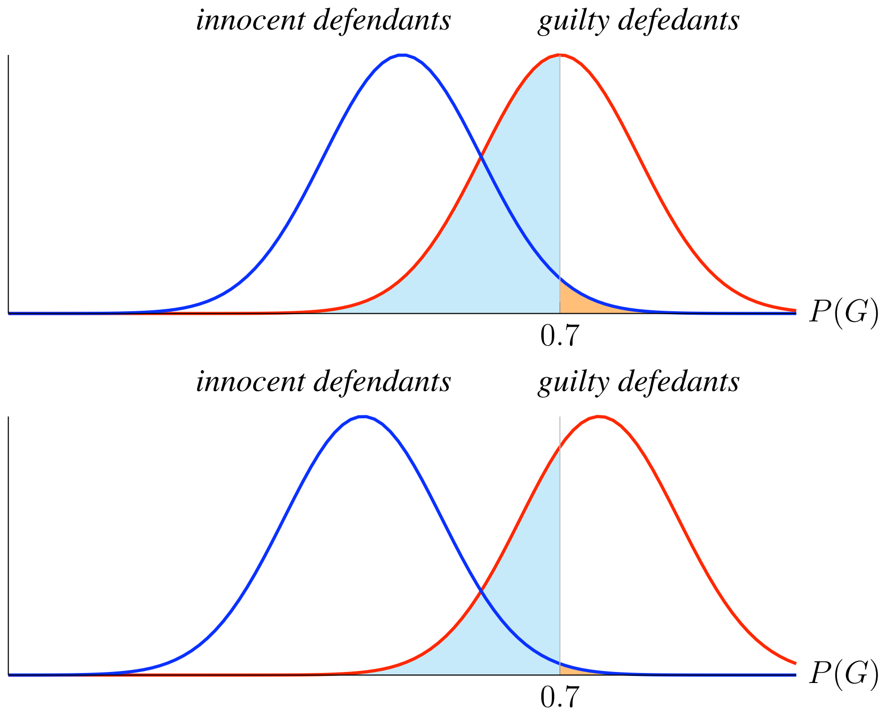
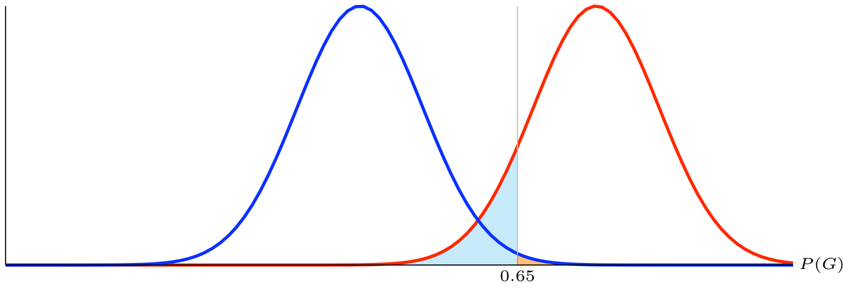


Figure 2.4: The difference between the two diagram is not the threshold line. The difference is in the distance between the averages of the two curves.

the two curves from each other, after all, represents an improvement of the trial system's ability to discriminate between innocent and guilty defendants.[21]

[21] Within SDT, we can also conceive of error reduction in another way. Instead of pushing the two curves further apart, the spread of the two curves can be reduced:



The spread has to do with *precision*. An analogy with a measurement instrument can be of help here. A measurement instrument, e.g. a scale, is fully precise whenever, for items of equal weight, it reports the

So, one way to reduce both *both* $P(C|I)$ and $P(A|G)$ is by ensuring that the trial system's discriminating power increases. (More on what this means, concretely, in the next section.) I should emphasize that the position of the threshold guilt probability need not affect the discriminating power of the system. Discriminating power and threshold, in other words, can be analytically defined independently of one another. In the SDT framework, the threshold is the position of the decision criterion, while the discriminating power is the distance between the averages of the two curves. The position of the decision criterion can change without affecting the distance between the averages of the two curves, and vice versa. This marks an important difference from error distribution and error reduction. Unlike threshold and discriminating power, we have seen that error distribution and error reduction are closely intertwined. Pursuing a desirable error distribution, in fact, can well clash with the goal of error reduction; this was clearly stated in the incompatibility claim.

The notion of discriminating power, I hold, allows us to work around the incompatibility claim. Instead of speaking of reducing errors as much as possible, we can speak of maximizing the trial system's discriminating power insofar as improving the system's discriminating power leads to error reduction. But a worry now arises. How can we be sure that no conflict will arise between the maximization of the trial system's discriminating power and pursuing a desirable error distribution? If discriminating power affects error reduction, and the latter can conflict with a desirable error distribution, the incompatibility claim might return in another form. The key here is that discriminating power does not fully control error reduction; discriminating power is not a replacement for error reduction. This is a tempting identification to make, but it is an incorrect one. Although improving the discriminating power of the trial system does positively affect error reduction, this does not mean that the error rate entirely depends on the system's discriminating power. Furthermore, let us compare the two diagrams both with a 0.7 threshold on page 59–59. The discriminating power is higher in the second diagram

same weight value. The less variability in the measurement results, the more precise the instrument. Similarly, a trial system which, for all the guilty defendants, gives widely different guilt probabilities, so that the guilt probabilities are significantly spread out, is less precise than a system whose guilt probabilities are more tightly concentrated around a certain probability value.

than it is in the first, and the error rate is consequently lower. The distribution of errors, however, is almost intact. The ratio of wrongful convictions to wrongful acquittals has not changed across the two diagrams. In both diagrams, the rate of wrongful convictions is significantly lower than the rate of wrongful acquittals. So, improving the discriminating power does not immediately affect error distribution, and consequently, it does not clash with the goal of promoting a desirable error distribution either. And note that this conclusion holds no matter how much we improve the system's discriminating power so long as the discriminating power does not reach perfection. So we have:

> COMPATIBILITY CLAIM. Maximizing a system's discriminating power (as a way to reduce errors) does not conflict with distributing errors in a desirable way (i.e. having proportionally fewer wrongful convictions relative to wrongful acquittals).

To be sure, the compatibility claim holds in the framework of Signal Detection Theory. It might be that, as a practical matter, maximizing the trial system's discriminating power will conflict with pursuing a desirable error distribution. I have not yet offered any guarantee that this will not happen in the context of criminal trials. For now, however, we should be content that the compatibility claim holds in theory.

With the compatibility claim in place, I am finally ready to rephrase step (c) in way that avoids the difficulties with (c*), as follows:

> (c-*qualified*) The rule of decision in criminal trials can only (or primarily) have the function of *distributing* errors in a desirable way *provided the trial system's discriminating power has been maximized.*

The qualified version of (c) inherits the intuitive and attractive feature of (c*) without being patently incoherent in its demand on the trial system. The new formulation is not patently incoherent because improving the trials system's discriminating power is a way to reduce errors, but in such a way that this does not openly clash with the goal of promoting a desirable error distribution.

***

The qualified version of step (c) calls for a qualified version of the threshold claim, i.e. the claim that an appropriately high threshold guilt should be the rule of decision in criminal trial. This is what I propose we use:

> THRESHOLD CLAIM-*qualified*. An appropriately high threshold guilt probability, say, 0.9 or 0.99, should be the decision criterion for criminal convictions *provided the trial system's discriminating power has been maximized.*

The qualified version of the threshold claim is not a refutation of legal probabilism. But it takes into account error reduction and discriminating power, besides error distribution and threshold guilt probability. This is my first refinement of legal probabilism, and I shall call it QUALIFIED PROBABILISM. The qualified version of probabilism makes it clear that the criminal standard of proof, understood as a probability threshold, does not obliterate the obligation to maximize the trial system's discriminating power and thus reduce errors. On the new formulation, the standard *qua* threshold comes into effect only as the last resort, so to speak, once any further improvement in the trial system's discriminating power is not possible.

I now argue that qualified probabilism is a first step toward reconciling traditionalists and probabilists. In sections 2.1 and 2.2 I offered a reconstruction of the dialectic between the two sides. I began with the probabilist threshold argument in support of the threshold claim, i.e. the claim that an appropriately high probability of guilt is enough to convict. Next, we saw the traditionalist random acquittal objection: if the motivation to set the threshold to a high value is that it promotes a desirable error distribution, the policy of randomly acquitting some defendants has also the effect of promoting a desirable error distribution, yet this does not mean we should endorse such a policy. This objection puts pressure on the threshold argument and the threshold claim. Next, we saw that when the traditionalists are asked to give their own decision criterion, they are likely to invoke something like Tribe's humanly attainable certainty. In response, the probabilists challenged Tribe's certainty by means of the stopping objection: humanely attainable certainty as the rule of decision in criminal trials cannot be any different from an appropriately high threshold guilt probability. All these opposing arguments are seemingly persuasive, but we cannot concurrently maintain all of them. I will now argue that by endorsing qualified probabilism and a qualified version of the threshold claim, we can accomodate the probabilist threshold argument and Tribe's demand for certainty. We can also address both the random acquittal objection and the stopping objection. Let me explain.

For one, qualified probabilism does not ignore the importance of setting a threshold guilt probability for the purpose of error distribution, so it satisfies (most of) the threshold argument. At the same time, qualified probabilism does not fall prey to the random acquittal objection outright. This objection shows that though randomly acquitting some defendants can promote a desirable error distribution, it should not be a policy that can be pursued. This is in tension with the probabilist threshold argument, but qualified probabilism can escape the objection. If some defendants were randomly acquitted, this would conflict with qualified probabilism because it would mean that the trial system openly disregarded the obligation to maximize its discriminating power.

Finally, let's see whether qualified probabilism can satisfy Tribe's demand for certainty while avoiding the stopping objection. If you recall, the stopping objection went as follows. Let's grant—as Tribe and other traditionalists want to suggest—that the trial system should strive for certainty, not just mere probability of guilt. Yet, the probabilists reasoned, the striving must stop at some point, and the criterion for when to stop can be either full certainty or a sufficiently high threshold guilt probability. Since the first criterion is unworkable, the probabilists concluded, a demand for certainty can be no different than a threshold guilt probability. Now, the probabilists are wrongly assuming that 'certainty' must be a very high probability or even a probability of one. But this need not be. There is another, and more sensible, interpretation of Tribe's demand for certainty, namely that the trial's system discriminating power should be maximized. Certainty, in this sense, would be closely tied with error reduction so long as discriminating power can positively affect error reduction. On the suggested interpretation, qualified probabilism (and the qualified version of the the threshold claim) would satisfy Tribe's demand for certainty because of the qualification 'provided the trial system's discriminating power has been maximized.'

And if we interpret Tribe's demand for certainty as a demand for maximizing the system's discriminating power, the stopping objection loses its bite. Of course, a limit will have to be set beyond which the system's discriminating power cannot be (expected to be) further improved; without such a limit, the trial system would be paralyzed. Now, this putative limit need not take the guise of a threshold guilt probability. It can take the guise of an *appropriately high threshold on the system's discriminating power*. A threshold guilt probability and discriminating power, after all, are distinct notions; they cannot be superimposed. Against qualified probabilism, then, the stopping objection fails.

I conclude that qualified probabilism can indeed serve as a promising theory of the legal-fact finding process. It is a theory which satisfies both the threshold argument and Tribe's certainty, while avoiding the random acquittal objection and the stopping objection. But what we have so far is only the skeleton of a theory. Besides a purely theoretical explanation, we are still in darkness as to what it means to improve the discriminating power of the trial system. Two open questions then arise:

> *Open Question 1:* What does it mean to reduce errors by means of improving the trial system's discriminating power? Which evidentiary rules or measures can reduce the system's error rate by improving its discriminating power?
>
> *Open question 2:* What does it mean to maximize the trial system's discriminating power? In other words, how can we define an appropriately high threshold on the system's discriminating power?

A third open question concerns whether the criminal standard of proof is in charge of error reduction or error distribution. If you recall, step (c) in the threshold argument asserts that the criminal standard's only (or primary) function is error distribution. On this score, qualified probabilism leaves open the possibility that the criminal standard be equated to a mere threshold probability serving the function of error distribution, while other evidentiary or procedural rules—e.g. rules of procedures, rules of weight—would be in charge of error reduction. This give rise to a third question:

> *Open Question 3:* Is the criminal standard in charge of error reduction, error distribution, or both?

Contrary to step (c) of the threshold argument, I will argue that the criminal standard, at least according to a plausible interpretation of it, imposes a demand on the criminal justice system for the purpose of reducing errors as well as distributing errors appropriately. I will also offer some remarks in response to the other open questions.

2.4 Reasonable doubts

Before seeing the connection between error reduction, discriminating power, and the criminal standard of proof, we need to spend some time on the standard itself. A common formulation of it is the expression 'guilt beyond a reasonable doubt.' But when is a doubt about a defendant's guilt reasonable or unreasonable? We are entering murky territory. Ask a judge or a lawyer, and you might hear something like this: beyond a reasonable doubt does not mean beyond any possible doubt; it means beyond any plausible, realistic, non-contrived doubt. It is not clear whether these paraphrases advance or obscure our intuitive understanding. Laudan (2006) has documented the disagreements among legal practitioners about the meaning of the criminal standard of proof. Perhaps wisely, the U.S. Supreme Court in Holland v. U.S. (1954) discouraged any definition of the criminal standard of proof, because, as the Court put it, 'attempts to explain the term "reasonable doubt" do not result in making it any clearer' (348 U.S. 121, 140). This is disturbing: we cannot define the standard, but we still send people to jail and even to death! Understandably, Dershowitz (1997) has called the Court's restraint 'an act of abject intellectual cowardice' (p. 69).

The reader should expect no systematic account of the criminal standard of proof here. More modestly, I will identify *four sources* which can give rise to reasonable doubts. The *first source* can be identified following the probabilist lead: a reasonable doubt about the defedant's guilt can arise whenever the evidence available at trial fails to meet a certain threshold guilt probability, e.g. 0.95 or higher. A reasonable doubt, on this interpretation, has to do with the (probabilistic) weaknesses of the supporting evidence. If the evidence fails to establish guilt with a probability as high as, say, 0.95, then a reasonable doubt would arise.

The *second source* has to do with the absence of evidence. The Supreme Court of Canada in R. v. Lifchus (1997) is adamant on this point. It writes that a reasonable doubt is also 'logically connected to the absence of evidence.' It is tempting to say that the second source has to do with the "incompleteness" of the evidence presented at trial. But this would be an exaggeration. It is an inherent feature of fallible evidence that it is incomplete, partial, fragmentary; there *could* always be more evidence. A better way to understand the second source is by reference to a *reasonable fact-finder's expectations about the evidence*. David Kaye (1986b) writes that 'a jury will expect to hear certain items of evidence in certain cases, and it may regard the failure to produce such evidence with devastating skepticism' (p. 663). For instance, in a drunk driving case, the fact-finders will expect testimony about the defendant's alcohol level and the use of a breathalyzer; if no mention of that is made, the fact-finders will be perplexed. A reasonable doubt about the defendant's guilt, then, can arise whenever the prosecutor fails to introduce items of evidence which he would normally, or reasonably, be expected to introduce.[22]

A *third source* has to do with what we might call the lack of resiliency—against challenges and objections—of the prosecutor's case. To see what this means, first note that some jury instructions in the past stressed the importance of minimizing the jurors' future regret.[23] This hints at the idea that a conviction is established beyond a reasonable doubt *only when* it is stable and not subject to future revision, rethinking, or retracting. Let's call this feature of a conviction its *resiliency*. A conviction is resilient whenever the discovery of future evidence, or the exercise of further scrutiny on the decision,

[22] There is some circularity here. In spelling out the second source of a reasonable doubt, I spoke of a *reasonable fact-finder's expectations about the evidence*. We made some progress, though: the unqualified concept 'reasonable doubt' has been qualified in terms of a reasonable fact-finder's expectations about the evidence.

[23] For instance, while giving instructions about the meaning of the criminal standard, Justice Lemuel Shaw's told the jurors they could convict only if they were 'confident that [they] will not be nagged by doubts about the propriety of [the] decision.' He added: 'make a decision you can live with.' Commonwealth v. Webster (1890). 59 Mass. (5 Cush.) 295, 320.

would not change the outcome. No conviction, however, can be definitely unshakable: in every criminal case, new and surprising evidence, or new and surprising ways to look at the evidence, could *always* change one's judgment about the case. The resiliency of a conviction, then, should be understood in a limited, relativized way. Let's call this *legal resiliency*. A conviction's legal resiliency can be relativized to trial proceedings and to the adversarial process—to what we might call *defense scrutiny.* This is a process by which the defense tests whether the prosecutor's case can withstand counter-evidence, challenges, objections, alternative hypotheses.[24] Legal resiliency, in this sense, would be satisfied whenever (i) the defense conducted a vigourous and thorough scrutiny of the prosecutor's case, and (ii) the prosecutor's case withstood the scrutiny. Conversely, legal resiliency would fail whenever conditions (i) or (ii) failed.[25] Accordingly, a reasonable doubt, if it results from lack of resiliency, can arise whenever the prosecutor's case was shaken by the defense scrutiny, or whenever the defense scrutiny did not take place or was not vigorous enough (for lack of resources, lack of procedural guarantees, etc.).

There is a *fourth source* of reasonable doubts. It is plain that at trial the prosecutor should establish the defendant's guilt, but there are two ways to interpret this. On one view, establishing guilt simply means establishing a number of propositions which jointly establish guilt; I call this the *atomistic view*. On another view, the prosecutor should proceed more holistically and offer an incriminating narrative, a cohesive reconstruction of what happened; I call this the *holistic view*. I think the holistic, narrative-based model more aptly describes what happens in a criminal trial. After all, it is difficult to establish guilt without offerring a narrative of guilt, especially because a crime involves the concomitant occurrence of the *actus reus* and the *mens rea* (Fletcher, 1998; Kaplan et al., 2008; Lippman, 2010). Spelling out the tight connection between the two is hard to do if one is trying to establish isolated propositions. To illustrate, suppose we

[24] I am not using the expression 'cross-examination' because this typically refers to the cross-examination of witnesses. Instead, scrutiny applies to the prosecutor's case as a whole, and it certainly includes cross-examination.

[25] Legal resiliency can fail in two cases: first, when the defense scrutiny shook the prosecutor's case against the defendant; second, when no proper scrutiny took place (e.g. because of bad lawyering). In other words, the prosecutor's case can fail to be resilient both *ex post* (if it could not withstand challenges) and *ex ante* (if it went unchallenged).

are in a murder case. Expert medical testimony shows that the victim died of poisoning and a search recovers a small quantity of poison in the defendant's house, the same type of poison which caused the victim's death. The two items of evidence suggest that the defendant was involved in the crime; but does this amount to guilt? The California Penal Code, sec. 189, for instance, is clear that, in order to establish the *mens rea* in a first-degree murder case, the prosecutor must show that the defendant 'manifested a deliberate intention unlawfully to take away the life of a fellow creature.' It seems difficult to do that without offerring a reasonably well-specified narrative of what happened before and after the killing. Further, when we know that the defendant possessed the poison that killed the victim, it is natural to ask additional questions. What if the defendant administered the poison by accident? What if the defendant actually wanted to kill himself and not the victim? What if the poison was placed in the defendant's house by a third party? These questions can only be answered by offerring a narrative of what happened before, during, and after the killing. Failing to offer satisfactory answers to these questions would give rise to reasonable doubts.

To recapitulate, I've suggested that there are four ways a reasonable doubt can arise (and there may be others): (1) when the guilt probability on the evidence is below a given threshold; (2) when the evidence is incomplete as far as a reasonable fact-finder's expectations about the evidence are concerned; (3) when the prosecutor's case against the defendant fails to be legally resilient; and finally, (4) when the incriminating narrative leaves unanswered some natural or reasonable questions about what happened. In other words, the criminal standard 'guilt beyond a reasonable doubt' imposes a fourfold requirement on what counts as a warranted conviction:

> EVIDENTIAL SUPPORT. The defendant's guilt probability on the evidence should meet an appropriately high threshold guilt probability.
>
> EVIDENTIAL COMPLETENESS. The evidence available at trial should be complete as far as a reasonable fact-finder's expectations are concerned.
>
> RESILIENCY. The prosecutor's case, based on the available evidence, should be legally resilient.
>
> NARRATIVITY. The narrative offered by the prosecutor should answer all

the natural or reasonable questions one may have about what happened.

***

What I have offered is merely the sketch of an account of the criminal standard of proof. I develop this account more extensively in chapter 7. What interests me for now is to add some concreteness to the abstract discussion of error distribution and error reduction in the previous section. So let's now examine evidential support, evidential completeness, resiliency, and narrativity from the point of view of error distribution and error reduction. Which function do they serve? By now there can be little doubt that the requirement of evidential support—understood as an appropriately high threshold guilt probability—serves the function of promoting a desirable error distribution. As the probabilist threshold argument shows, a demanding threshold guilt probability makes convictions harder, thereby distributing errors as to minimize wrongful convictions at the cost of increasing wrongful acquittals (see section 2.1). We also know that a higher threshold guilt probability has the effect of increasing the overall rate of errors (see section 2.3). So, evidential support serves the function of error distribution, but not that of error reduction.

Turning now to evidential completeness, resiliency, and narrativity, I will argue that they all contribute to improve the trial system's discriminating power, and in so doing, they serves the function of error reduction. Before I do so, I shall lay out a general rationale for how the system's discriminating power can be improved. If we consider the trial system as a whole, the application of evidentiary rules which tend to unilaterally and systematically favour the defense or the prosecution will not improve the system's discriminating power. Any one-sided measure is likely to make it easier to convict or acquit, without affecting the discriminating power. Such one-sided measures include: giving more resources to the prosecutor only (or to the defense only); giving the power to subpoena witnesses only to the prosecution (or only to the defense); etc. These measures are likely to make available more incriminating evidence than exculpatory evidence (or more exculpatory evidence than incriminating evidence). Yet, the discriminating power is that of discriminating *between* innocent and guilty defendants. In order to tell apart the guilty from the innocent, without having any form of prescience, the trial system should have the best possible evidence *from both sides*.[26] So any evidentiary measure should not be one-sided if it is to positively affect the discriminating power of the trial system and thus reduce the overall error rate. Here is a concrete example. The introduction of DNA evidence in criminal trials has represented a tremendous improvement in terms of discriminating power.[27] DNA evidence is not one-sided and it can potentially benefit both parties. DNA evidence can benefit the prosecutor (if it turns out that the genetic trace at the crime scene matches the defendant), but it can also benefit the defendant (if it turns out that the genetic trace at the crime does not match the defendant).[28]

If the foregoing is correct, a certain evidentiary measure improves the trial system's discriminating power provided the party who benefits from the evidentiary measure cannot be identified beforehand by a third party who knows nothing about the case; the facts of reality, so to say—the facts of guilt or innocence—should decide which party will benefit. I call this the *no predictable beneficiary criterion*. Whichever evidentiary

[26] This point is reminiscent of Mill's argument in *On Liberty*. He argues that the free and rational confrontation of different opinions is an effective way—the best we have—for the discovery of the truth. Things are not equally straightforward in criminal trials, however. Each party is more concerned with winning than with seeking the truth (Stuntz, 2011). As a prosecutor amasses evidence against the accused, he might do so for the sake of boosting his ego. As the defense introduces more evidence or raises challenges, it might introduce misleading evidence for the sake of winning. The presence of such uncooperative tendencies suggest that the criminal justice system should maximize the incentives for producing non-misleading evidence and for formulating non-misleading arguments. A game-theoretic analysis might be helpful here. For a game-theoretic defense of the right to silence, see (Seidmann and Stein, 2000).

[27] For a quick introduction to DNA evidence and its uses in the courtroom, see Wasserman (2008). For a more in-depth treatment, see Kaye and Sensabaugh (2000).

[28] When traces of blood, semen, saliva, skin tissue, etc. are found at the crime scene, laboratory analyses can create a DNA *profile* from the traces. A DNA profile is a codified representation of certain portions of the human genome which tend to be different across individuals. Once a profile is created from the traces, it is compared against a suspect's blood, semen, saliva, skin tissue, etc. from which another DNA profile is created. The purpose of the comparison is to find a genetic *match* between the two profiles. If a match is found, this would constitute strong yet not infallible evidence that the suspect is the source of the traces found at the crime scene. What makes DNA evidence particularly powerful is its statistical underpinnings. DNA profiles, albeit not unique, are highly discriminating because they each occur very rarely. The rarity of a DNA profile is expressed by a statistical frequency, sometimes as astronomically small as 1 in 50 billion, representing the profile's expected frequency in a population. See Kaye (2010b).

measure satisfies this criterion should positively contribute to discriminating power and thus to error reduction. We shall now see whether evidential completeness, resiliency, and narrativity—understood as systemic evidentiary measures that apply to each single trial—satisfy the no predictable beneficiary criterion. If they do, they would positively contribute to the trial system's discriminating power and thus to error reduction.

Let me begin with evidential completeness, i.e. the requirement that the evidence be as complete as can be reasonably expected. This requirement puts the burden of producing evidence on the prosecutor, and on the prosecutor only. It is the prosecutor who has to present as complete evidence as possible. But by placing a burden on the prosecutor, evidential completeness does not systematically favour the other party. Adding more evidence—or requiring that the evidence be as complete as possible—will favour the prosecutor in some cases and the defendant in other cases. The required items of evidence will be exculpatory in some cases and incriminating in other caes. Compliance with evidential completeness can cut both ways, so it satisfies the no predictable beneficiary criterion. The same can be said of legal resiliency. It is the responsibility of the defense to scrutinize the prosecutor's case in order to assess its resiliency against objections and challenges. But the outcome of the defense scrutiny cannot be predicted beforehand. The scrutiny can result in the weakening of the prosecutor's case (if it turns out that the case succumbs to the defense challenges) or in its strengthening (if it turns out that the prosecutor's case resists the challenges). Resiliency can cut both ways, and thus it satisfies the no predictable beneficiary criterion. To be sure, whenever resiliency is satisfied, this always goes in favour of the prosecutor's case. What can can cut both ways, then, is not resiliency itself, but rather the defense scrutiny which is required to assess the resiliency of the prosecutor's case.[29]

[29] Henry Wigmore famously said that cross-examination (defense scrutiny, in my terminology) is the "greatest legal engine for the discovery of truth." What he had in mind by 'truth,' I think, is what I call 'discriminating power.' Wigmore's claim has been quoted several times, but little research has been done to show that it in fact holds. See Davis v. Alaska, 415 U.S. 308 (1974) ['cross examination is the principal means by which the believability of a witness and the truth of his testimony are tested' *at* 316]; U.S. v. Salerno, 505 U.S. 317 (1992) ['in the Anglo American legal system cross examination is the principal means of undermining the credibility of a witness whose testimony is false or inaccurate']. As far as actual legal proceedings are concerned, the epistemic virtues of cross-examination have never been fully

What about narrativity? The requirement of narrativity is that the prosecutor should offer a reconstruction of the crime that is as complete, detailed, and specific as it can be reasonably expected. Another way to formulate the narrativity requirement is to say that the prosecutor's reconstruction of the crime should answer all the reasonable questions one might have about what happened. Just like evidential completeness and legal resiliency, the narrativity requirement can cut both ways. Suppose the prosecutor offers a reasonably detailed and specific narrative of the crime, thereby meeting the narrativity requirement. Now, as soon as a narrative asserts certain claims, it exposes itself to challenges relative to those claims. If a narrative says more, it exposes itself to more challenges.[30] So, the narrativity requirement forces the prosecutor's case to be exposed to a certain number of challenges. This will favour the prosecutor in some cases and the defense in other cases. If the prosecutor's narrative can withstand the challenges leveled by the defense, the prosecutor's case will come out stronger. If not, the prosecutor's case will come out weaker. Once again, compliance with the narrativity requirement can cut both ways and thus it satisfies the no predictable beneficiary criterion.

It is instructive to compare evidential completeness, resiliency, and narrativity with a high threshold guilt probability. Does a high threshold guilt probability—*qua* an evidentiary measure on the trial system—satisfy the no predictable beneficiary criterion? It seems not. A high threshold guilt probability makes convictions harder, and thus it systematically benefits defendants and not prosecutors. This is what a high threshold is supposed to do; it constitutes a bias of the system in favor of defendants. And interestingly enough, the fact that a high threshold guilt probability does not satisfy the criterion fits well with the claim from section 2.3 that a high threshold guilt does not promote the reduction of errors. All in all, I conclude that evidential completeness, resiliency (understood as defense scrutiny), and narrativity all positively contribute to discriminating power, because they all satisfy the no predictable beneficiary criterion.

demonstrated, although they haven't been disproven either. For a review of the literature and an epistemic defense of cross-examination, see e.g. (Sanchirico, 2009).

[30] The correlation between the specificity and level of detail of a narrative and its susceptibility to challenges echoes an idea of Karl Popper. He believed that the hallmark of scientific hypotheses is that they are falsifiable, i.e. susceptible to be falsified by further observations (Popper, 1935, 2002).

***

It is time to address the three open questions at the end of section 2.3. What does it mean to improve the trial system's discriminating power? What does it mean to maximize it? Is the criminal standard in charge of error reduction, error distribution, or both?

I have just argued that evidential completeness, resiliency, and narrativity all positively contribute to the trial system's discriminating power. So whenever each individual trial satisfies the three requirements, the system's discriminating power benefits. The question remains of what it means to maximize discriminating power. Maximization hints at the idea of a threshold. We've seen that such a threshold cannot take the form of a threshold guilt probability. But if evidential completeness, resiliency (understood as defense scrutiny), and narrativity all positively contribute to discriminating power, the threshold in question can be broken down into three items: that the evidence be as complete as it can be reasonably expected; that the prosecutor's case be vigorously scrutinized by the defense; and finally, that the prosecutor's narrative be reasonably specific so as to answer all the natural questions a reasonable fact-finder might have about what happened. These three requirements (or thresholds) are linked to discriminating power as follows:

> DISCRIMINATING POWER THRESHOLD. The trial system's discriminating power has been maximized whenever the requirements of evidential completeness, legal resiliency (understood as defense scrutiny), and narrativity are systematically complied with.

To be sure, there might be other factors that contribute to discriminating power. So the above must be understood as a close approximation of what its means to maximize the system's discriminating power.

I now turn to the third question. Is the criminal standard in charge of error reduction, error distribution, or both? My descriptive analysis of the formula 'beyond a reasonable doubt' suggested that evidential completeness, resiliency, and narrativity are part of the criminal standard. And since, as I've shown, the latter contribute to improving the system's discriminating power and thus to reducing errors, so must the criminal standard.

The standard, of course, can *also* serve the function of error distribution so long as it includes a threshold guilt probability.[31] The account of the criminal standard of proof I have in mind is this. A conviction is warranted whenever (1) the probability of guilt meets an appropriately high threshold; (2) the evidence is reasonably complete; (3) the prosecutor's case is legally resiliency; (4) and the prosecutor's narrative is reasonably specific. This is what I earlier called the fourfold requirement. The question I address in the next section is whether the legal probabilists have the resources to work out a probabilistic account of the criminal standard along the lines of the fourfold requirement. My answer will be affirmative.

2.5 PROBABILISM—MORE AND BETTER

In giving an account of the criminal standard of proof, the legal probabilists pay exclusive attention to the probability of guilt. Instead, if my analysis of the formula 'guilt beyond a reasonable doubt' is on the right track, evidential completeness, resiliency, and narrativity should also be taken into consideration. But a problem for probabilism now arises. Legal probabilism, as a conceptual framework to understand the criminal standard of proof, contains only two key notions: guilt probability on the evidence, and a threshold guilt probability. The first is a measure of the strength of the supporting evidence; the second is a decision criterion for criminal convictions. It is not immediately clear how evidential completeness, resiliency, and narrativity can be accommodated within so parsimonious a framework.[32]

[31] I have not addressed the normative question. What *should* the proper function of the criminal standard be? Should it be error reduction, a desirable error distribution, or both? I think that a trial as a whole should promote both error reduction and a desirable error distribution. Now, when it comes to the final decision, the decision criterion—i.e. the criminal standard of proof—should plausibly serve as the final check on whether the trial as a whole has been working properly. This means that the standard should make sure that errors have been effectively kept to a minimum and that they are distributed in a desirable way. Even at the normative level, it seems, the standard's proper function should be error reduction as well as error distribution.

[32] The probabilists can insist that guilt probability is *also* a measure of evidential completeness, resiliency, and narrativity. Let's grant the probabilists, for the sake of argument, that this move is feasible. Guilt probability would then become a "catch-all" notion; we can denote it by *guilt probability** so as

The probabilists can enrich their framework by adding probability-based accounts of evidential completeness, resiliency, and narrativity. To my knowledge, no legal probabilist has undertaken the task in any systematic way. It is plausible to think, however, that the probabilist framework can be suitably extended without denaturing it. In this section, I will outline how this can be done. Let's begin with evidential completeness. There are good reasons to think that evidential completeness is an implicit requirement of any probability-based account of evidence. In Confirmation Theory, for instance, the requirement of evidential completeness often takes the form of the *total evidence requirement*. The latter (roughly) amounts to the idea that the evidence confirming one's hypothesis should be one's total evidence. One way to capture the total (or complete) evidence requirement in probabilistic terms is by means of second-order probabilities. Consider the probabilistic claim $Pr(H|E) = r$, where H is some proposition and E is the supporting evidence. If the evidence E is truly *all* the evidence, we can assign a second-order probability close to one hundred perent to the probabilistic claim $Pr(H|E) = r$.

Let's now turn to legal resiliency. First of all, probability-minded epistemologists well admit of a difference between 'high probability' and 'resiliently high probability' (Skyrms, 1977, 1980; Leitgeb, 2010). Suppose the probability that it rained, given that the road is now wet, equals 0.9, because e.g. experience tells us that, in 1 out of 10 cases,

to avoid mistaking it for the standard *guilt probability*. Invoking a catch-all notion could be the solution, yet it is also where the difficulty lies. By embedding different notions into one, legal probabilism, instead of clarifying the logic of legal fact-finding, might become hopelessly confused. Let me explain. We have seen that a threshold guilt probability controls error distribution, while evidential completeness, resiliency, and narrativity control error reduction. A catch-all threshold guilt probability* would instead control both error distribution and error reduction. The consequence of this, I think, is that the initial motivation for equating the criminal standard of proof to a high threshold probability* will be undermined. To see why, recall that a high threshold is intended to promote a certain distribution of wrongful convictions and wrongful acquittals, a distribution that agrees with the widely held belief that a wrongful conviction is more harmful than a wrongful acquittal. But if guilt probability* controls error reduction, and not only error distribution, it is no longer clear where the new threshold should be located. Should it be more, less, or equally demanding? How would the threshold affect error distribution given that, at the same time, it also affects the reduction of errors? The shift from guilt probability to guilt probability* requires one to rethink the argument that equates the standard of proof to a high threshold guilt probability*. I conclude that it would be unwise for the probabilists—and possibly a source of unnecessary confusion—to pretend that guilt probability* can be a catch-all notion.

a wet road follows the rain. Though 0.9 is a high probability, you might wonder whether it can change significantly in light of future evidence or alternative explanations. For instance, consider the hypothesis that the sprinklers have been watering the grass on the sidewalks. This hypothesis, if true, would bring down the value of the initial probability assignment. The probability than it rained, given that road is now wet and—*in addition*—that the sprinklers have been watering the grass on the sidewalks, must be significantly lower than the initial 0.9 value. So, this value was high but not very resilient. Note that the resiliency of a probability value assignment is always relative to a proposition (or a set of propositions) describing an item of evidence or an hypothesis. In our example, the value 0.9 was not very resilient relative to the proposition 'the sprinklers have been watering the grass on the sidewalks.'[33]

The notion of resiliency thus described can be applied—*mutatis mutandis*—to the trial context. The legal probabilists can offer an account of the resiliency of probabilistic claims of the form $P(G|E) = r$, where G stands for the defendant's guilt and E denotes the available evidence (at some point during the trial). In criminal trials, the set of propositions relative to which the probability of guilt can vary—that is, the set of propositions relative to which resiliency is measured—will consist of the challenges,

[33] Let's state all this more generally. Suppose the probability of a proposition A, given a body of evidence E, equals some value r between 0 and 1; we write $P(A|E) = r$. The resiliency of a statement that assigns a probability value to a proposition (a probability statement, for short) can be formulated as the complement of the statement's variability, so that the more variability, the less resiliency. Now, the variability of $P(A|E) = r$, relative to a proposition π_i, is the absolute difference between r and the probability of A given E *and* π_i, or in short, $|r - P(A|E \wedge \pi_i)|$. Variability can be defined more generally over a set of propositions $\Pi = \{\pi_1, \pi_2, \pi_3 \dots\}$ by taking the widest variability, namely $max_i\{|r - P(A|E \wedge \pi_i)|\}$, where π_i is any proposition in Π such that π_i and E are consistent. (E and π_i must be consistent because the conditional probability of a proposition given an inconsistent proposition is typically undefined.) Following Skyrms (1980), a definition of resiliency can now be stated:

> The SKYRMS RESILIENCY of a conditional probability statement $P(A|E) = r$, relative to a set of propositions Π, is given by 1 *minus* its variability, i.e. $1 - max_i\{|r - P(A|E \wedge \pi_i)|\}$. (If the variability of a conditional probability statement is zero, then its Skyrms resiliency will be complete, as expected.)

The Skyrms resiliency of a probability statement comes in degrees between 0 and 1, and it is always relative to a set of propositions Π. Full and absolute resiliency—i.e. resiliency with degree one relative to all the propositions in a language—is very hard to come by.

objections, and counter-evidence which the defense presented (or could have presented) during the scrutiny of the prosecutor's case.[34]

Finally, let's consider narrativity. The probabilists can certainly be open to a more holistic approach, one in which establishing guilt does not amont to establishing individual propositions, but rather, it amounts to establishing a comprehensive and coherent narrative of guilt. The probabilists would need to offer a well-thought account of what it means for a narrative to be more or less specific and of what it means for different claims in a narrative to cohere with one another or to corroborate one another. Since such a well-thought account is yet to be given in the literature, the legal probabilists can hardly be faulted for not having one.

In conclusion, I've sketched how the probabilists might go about offerring a probability-based account of evidential completeness, resiliency, and narrativity. More needs to be said. My remarks here were primarily meant to show that legal probabilism can be extended beyond the mere concept of a threshold guilt probability. We can now make a plausible supposition. Suppose that the legal probabilists manage to formulate a probability-based metric for evidential completeness, legal resiliency, and narrativity. They would then be in a position to offer a more sophisticated account of the criminal standard of proof, more sophisticated than a mere high threshold guilt probability. This account would consist of four probability-based thresholds: a guilt probability threshold; an evidential completeness threshold; a resiliency threshold; and a narrativity threshold. I call this NEW PROBABILISM.

[34] Let Σ denote the set of propositions describing challenges, objections, and counter-evidence. The propositions in Σ will be of two types: items of exculpatory evidence and exculpatory hypotheses about what happened. Now, Skyrms resiliency—see previous footnote—does not distinguish between upward or downward variability, since it uses the absolute value of the difference. In the context of criminal trials, however, we are mostly interested in downward variations of the guilt probability on the evidence. The problematic case arises when the guilt probability goes *down*, maybe even below the threshold guilt probability. Legal resiliency, then, is a subcase of Skyrms resiliency, as follows:

> The LEGAL RESILIENCY of a conditional probability statement $P(G|E) = r$, relative to a set of propositions Σ, is given by 1 *minus* its downward variability, i.e. $1 - max_i\{|r - P(A|E \wedge \pi_i)|\}$ restricted to only the π_i's such that $r \geq P(A|E \wedge \pi_i)$.

2.6 Reconciliation?

I began this chapter by promising that I would reconcile probabilists and traditionalists. New probabilism is my attempt at a reconciliation. For one, new probabilism should satisfy the open-minded probabilists; but will it appease the traditionalists? This is more difficult to say. I begin by examining some of the complaints about legal probabilism and see if they apply to new probabilism.

> Moral complaint. Criminal trials are not about gambling with people's lives; the criteria for a conviction cannot be based on long-run, expected ratios of errors; each defendant is unique and should be respected as such (Wasserman, 1991; Pundik, 2009).
>
> Procedural complaint. Trial proceedings are not monological probability assignments; they are a back-and-forth between two opposing parties; cross-examination plays a crucial role (Stein, 2005).
>
> Reasoning-based complaint. Legal reasoning is not merely probabilistic: it is not only evidence-to-hypothesis reasoning (i.e. reasoning about how well the evidence *supports* the hypothesis), but also hypothesis-to-evidence reasoning (reasoning about how well the hypothesis *explains* the evidence) (Wells, 1992); legal reasoning is best modeled by inference to the best of explanation (Dant, 1988; Pardo and Allen, 2008); legal reasoning is coherence-based (Haack, 2011); legal reasoning is analogical (Picinali, 2012).

I think that new probabilism does not fall prey to these complaints. First of all, in new probabilism, promoting a desirable long-run ratio of errors is not the sole function of the criminal standard. The criminal standard also plays a pivotal role in error reduction. Evidential completeness, legal resiliency, and narrativity promote error reduction by ensuring that as much evidence as possible be gathered; that an extensive scrutiny of the prosecutor's case be conducted; and that the reconstruction of the crime, offered by the prosecutor, be as specific as possible. The importance of error reduction in new probabilism can well be rooted in the recognition of the dignity and individuality of each defendant. This should alleviate the moral complaint. In order to address the procedural complaint, it is enough to note that legal resiliency—as an integral part of the criminal standard according to new probabilism—is closely linked with the defense scrutiny and with the process of cross-examination.

Finally, the reasoning-based complaint is the most difficult to address. It is premised on a variety of non-probability based accounts of legal reasoning which would be impossible to discuss in detail here. One thing I can say, though, is that new probabilism does not view legal reasoning as mere evidence-to-hypothesis probabilistic reasoning. Think of narrativity and of how gaps or oddities in the prosecutor's narrative might trigger reasonable questions about what happened, questions which are in need of an answer. What is going on here is not merely evidence-to-hypothesis probabilistic reasoning. Reasonable questions about what happened, after all, arise because the narrative *as is* fails to adequately explain the available evidence. Think also of legal resiliency. If you recall, the set of propositions relative to which legal resiliency is measured, contains alternative hypotheses about what happened. Legal resiliency is sensitive to the possibility that if these hypotheses were true, they could change the initial probability of guilt. The notion of legal resiliency, then, exploits a form of hypothetical reasoning which is not merely evidence-to-hypothesis probabilistic reasoning.

***

The review of the traditionalist complaints suggests that new probabilism does not fall prey to them, or least, not as straightforwardly as the old, rough version of probabilism. New probabilism is a middle-ground position which both traditionalists and probabilists can endorse. Let me conclude by discussing whether the traditionalists might still justifiably resist new probabilism.

In new probabilism, the criminal standard of proof is viewed as a fourfold threshold: for guilt probability; for evidential completeness; for legal resiliency; and for narrativity. Unlike the old probabilist version, the fourfold account is able to control both error distribution and error reduction insofar as guilt probability controls the former and evidential completeness, resiliency, and narrativity the latter. Let's suppose that, as a device for controlling error reduction, the criminal standard has been tuned—through the synergic contribution of evidential completeness, resiliency, and narrativity—to aim at 1 percent overall rate of errors, inclusive of both wrongful convictions and wrongful acquittals.

The traditionalists might still resist new probabilism. They might observe—as Lawrence Tribe has done—that *a system that explicitly aims at making no more than 1 percent overall rate of errors is much different from (and much worse than) a system that, while aiming at making no mistake whatsoever, ends up with, say, a 1 percent overall rate of errors*.[35] Some traditionalists might tolerate the latter but not the former system, which is the one, presumably, resulting from endorsing new probabilism. The difference can also be put in terms of *intended errors*: although the *de facto* error rate might be the same in the two systems, in one system the errors are intended, while in the other they are not. And if so, one system does look worse than the other.

We should be careful here. New probabilism, as I see it, is a development of qualified probabilism from section 2.3. Part of qualified probabilism is the idea that the trial system should strive to maximize its discriminating power for the purpose of reducing errors. Declaring a fourfold threshold is a way to spell out what it means to do one's best to reduce errors (by means of discriminating power maximization), on the tacit assumption that doing one's best has a limit. By declaring a fourfold threshold, new probabilism is not openly aiming at a certain error rate (in the same way in which letting a friend die, when nothing else can be done, does not mean to aim at killing a friend). New probabilism, however, is openly acknowledging—and possibly predicting—that there will be a certain error rate. And this is the truth of probabilism we cannot deny, whether we like it or not. I wish to make this point more vidid. Consider this scenario:

> ALMOST PARADISE. Imagine a trial in which the incriminating evidence collected is as complete as possible; the prosecutor's case is scrutinized as

[35] Lawrence Tribe makes a similar point. He writes:

> Tolerating a system in which perhaps one innocent man in a hundred is erroneously convicted despite each jury's attempt to make *as few mistakes as possible* is in this respect vastly different from instructing a jury to *aim at a 1 percent rate* . . . of mistaken convictions.' Tribe (1971b), footnote 143, p. 1347.

> vigorously as possible and it has successfully resisted such a scrutiny; the prosecutor's narrative is as precise and specific about what happened as possible. After all that, the result is an estimate of the defendant's probability of guilt. The guilt probability meets a sufficiently high threshold, e.g. 0.99 or higher. Given the high probability of guilt, the trial moves forward and the defendant is convicted.

The conviction described in *Almost Paradise* meets the fourfold threshold of new probabilism. In the trial, after all, the participants have done their best to minimize errors. The resulting conviction, however, is preceded by a quantified statement of the possibility of error, in terms of a probability of guilt which falls short of one. The quantified statement of the possibility of error constitutes an open admission that some errors are inevitable. We can do nothing about that. New probabilism, just like old probabilism, does not ignore this inevitability and it makes it a factor that enters the decision making process.

What ground would the traditionalists have for resisting the conviction in *Almost Paradise*? I see no ground for that. The defendant's probability of guilt is high; the incriminating evidence is as complete as possible; the prosecutor's cases has been scrutinized as thoroughly as possible; and the prosecutor's narrative is as detailed as possible. If in such circumstances a conviction were still unwarranted, I do not see what could make it warranted except the absolute certainty of guilt. But the latter would be an irrational expectation to have—one which would constantly paralyze that trial system. I conclude that the traditionalists have no principled reason to resist a conviction in *Almost Paradise*, and thus have no principled reason to reject new probabilism.

2.7 ADDENDUM: EPISTEMIC LUCK AND ERROR REDUCTION

As seen in chapter 1, the moral philosopher Judith Thomson thinks that an item of evidence counts as individualized provided it is casually connected, in the appropriate way, to the event or state of affairs that the evidence is purported to establish. On her account, an eyewitness identification would be individualized evidence because, presumably, the witness was, at some point in time, causally connected through her sensory apparatus with the person being identified as the perpetrator. Thomson formulates a two-pronged requirement for a conviction. For her, a criminal conviction is acceptable provided (i) the probability of guilt on the evidence is sufficiently high, and (ii) the supporting evidence is individualized, i.e. it is causally connected, in the appropriate way, with the disputed event. Part (i) is inherited from the probabilists, while part (ii) is Thomson's. Note that Thomson's two-pronged requirement is an answer to the stopping objection because it is different from the probabilist idea that a conviction should simply meet an appropriately high threshold guilt probability, and also, it is different from a demand for unattainable certainty.

One lingering question is, why should the evidence be causally connected with the event of the crime? Thomson thinks that otherwise it would be a matter of luck whether a verdict ends up being right. (We are here dealing with *epistemic luck*, not moral luck. Epistemic luck is the type of luck that derives from being accidentally right about a certain factual question e.g. whether the accused is guilty or innocent.) The case Thomson has in mind is one in which a guilty verdict rests on the mere statistical chances that a defendant is guilty. Think about the prison yard scenario from chapter 1. In that scenario, the defendant Eschaton is on trial because the statistics-based probability that he killed the prison guard equals 99:100. Intuitively, it would be a matter of (epistemic) luck if the jury convicts, and Eschaton is in fact guilty. Thomson thinks that, in contrast, if the jury relied on a causally-connected item of evidence pointing toward Eschaton's guilt, the jury would be shielded from epistemic luck.

***

Thomson's argument is a step forward from Tribe's demand for certainty. Unlike Tribe, Thomson avoided introducing a vaguely specified demand for certainty, but she instead argued that an acceptable conviction should *not be based on luck* (or more precisely, that the correctness of a conviction should not be a matter of epistemic luck). The problem is that, at least for the probabilists, Thomson's claim will sound the same as the claim that a conviction should be *infallible* or *certain*. After all, the difference between a luck-free conviction and an infallible or certain conviction is not obvious. And since the latter is impossible, the probabilists have a *prima facie argument* that the former must also be impossible.

Thomson might clarify the matter by saying that a demand for a luck-free conviction does not amount to the much stronger, and admittedly unrealistic, demand for certainty and infallibility. How so? For Thomson, a conviction based on a piece of evidence that is causally connected with the contested event of the crime would be luck-free, yet it would not be infallible because a causal connection *might* have failed. Recall that, for Thomson, a conviction is acceptable provided (i) the probability of guilt on the evidence is sufficiently high, and (ii) the supporting evidence is individualized, i.e. it is causally connected, in the appropriate way, with the disputed event. Neither clause (i) nor (ii) amounts to a demand for certainty. But the probabilists might rebut:

> INFALLIBILITY/NO-LUCK OBJECTION. Let's consider the causal connection of the evidence with the disputed event. Now, either such a connection *might* fail or it is thought to be infallible; if the first option obtains, then luck would not be eliminated, because whether the connection fails or not would still be a matter of luck; if the second option obtains, we would—unrealistically—deny the fallibility of human evidence.

The probabilists can dismiss Thomson's argument because they postulate an immediate connection between fallibility and luck. In other words, they think that if a piece of evidence is *fallible*, e.g. it supports a conclusion with a probability short of certainty, then believing the conclusion on the basis of the (fallible) evidence and being right about the conclusion would automatically be a matter of epistemic *luck*. For the probabilists, any correct belief based on fallibile evidence is epistemically lucky. Thomson, and the traditionalists more generally, should instead be able to resist the immediate connection between fallibility and luck.

Thomson thinks that a verdict based on causally-connected evidence would *not* be correct as a matter of luck. She thinks that causality trumps epistemic luck, but she offers no clear reason for thinking so. And here is where the no-luck/infallibility objection arises. Despite the putative causal connection, isn't a verdict based on a eyewitness evidence also prone to be correct as a matter of epistemic luck unless we require the witness to be infallible? This is the infallibility/no-luck objection. In what follows, I will address the objection by connecting the notion of error reduction which I developed in this chapter with Thomson's idea of a luck-free verdict.

***

Causal connections evoke the idea of control, and so long as we have control of a situation, its outcome should not be a matter of luck. Consider now this scenario:

> DRIVING CAREFULLY. I drive my car down the highway at night. I am awake and drive carefully. My car could break down; I could get into an accident; anything could happen. But none of these things happen to me, and I get gome safely.

An accident could have happened, but we would not say I was lucky I got home. Or would we? Consider a second scenario:

> DRIVING BLINDLY. I am driving at night with my eyes closed. The highway is straight. Reliable, well-tested statistics tell me that virtually no car is on the road at that time. There is only a 1 in 100,000 chance to find one. As it turns out, I get home safely. (I opened my eyes only at the very end, to exit the highway and park.)

I was lucky I did not get into an accident, wasn't I? It is (practical) luck that I got home safe after driving with my eyes closed. Keeping my eyes open would have given me *more control*, no matter what the background chances of hitting a car were. When I decided to drive with my eyes closed, I gave up a degree of control I would normally enjoy. We can thus identify two types of luck:

> ENDEMIC LUCK. The occurrence of a positive outcome is beyond our control *and we cannot do anything about it*.
>
> ELIMINABLE LUCK. The occurrence of a positive outcome is beyond our control *because we openly surrendered our control over it*.[36]

Endemic luck has to do with the non-occurrence of, for example, an unpredictable failing of our senses, a mechanical car problem, an error in our measurement instruments, a failure of "the system." Beyond a certain point, we can simply do nothing about these incidents. We are lucky (in the endemic sense) if nothing bad happens to us. Eliminable luck, instead, is within what we can control, in the sense that we can shield ourselves from its influence. If I close my eyes while driving instead of keeping them open, I am openly lowering my control of a situation on which I could exercise better control. This is what gamblers do. They openly decide that the outcome of a coin toss will decide where their money should go; they purposefully surrender their control.

Let us now relate all this to epistemic luck. A parellel with a scenario that is familiar to epistemologists is helpful here. Suppose I bought a lottery ticket, knowing that my chances of losing were exorbitantly high. Would it make sense for me to say: oh well, I'll just tear up my lottery ticket? It wouldn't. Why is that? Why shouldn't I conclude that I lost (or will lose), given the statistical odds of an almost sure loss? As I am holding a lottery ticket, I can expect to acquire more evidence relatively soon, evidence that together with the statistics would constitute better evidence overall. Given the normal, expected path of evidence acquisition, a newspaper or the TV will tell me the outcome of the lottery drawing. Simply ignoring these potential sources of further information,

[36] Tort law makes a distinction between 'driving through an intersection with the individualized awareness that we might hit a pedestrian' and 'starting a construction project with the statistical awareness that some workers might get injured.' Injuring others is regarded as more or less culpable depending on whether one acted under individualized knowledge of risk or under statistical knowledge of risk. It is an interesting question why that should be so (Simons, 2011). My sense is that the tort law distinction between 'acting under individualized knowledge of risk' and 'acting under statistical knowledge of risk' can also be read as a distinction between 'acting under eliminable luck or risk' and 'acting under endemic luck or risk.'

and just be content with the statistics, is just like closing one's eyes while driving. In the highway case, we should not give up control of the situation; it is a control which we normally have. Similarly, in the lottery example, we should not draw a conclusion before we become acquainted with additional evidence besides the statistics; and this is additional evidence which we expect to acquire in normal circumstances. The upshot is that we would justifiably feel uneasy were we to conduct our lives by openly surrendering our control over what we are normally able to control. Similarly, we would justifiably feel uneasy were we to draw conclusions from a body of evidence by openly ignoring further evidence which we are normally able to acquire. (More on this in chapter 6.)

How do these observations apply to criminal trials? In criminal trials, too, jurors should not pronounce a verdict until the prosecution and defense have done their best to present all available evidence and have done their best to spot weaknesses in the case of the opposing party. In cases like the prison yard scenario, it seems that the prosecutor has been lazy, to put it bluntly. If Eschaton were convicted, it would be a matter of epistemic luck—the eliminable kind of luck—if he was, in fact, guilty. The reason is that both the prosecution and the defense should have done more; they should have offered more evidence and a more precise reconstruction of what happened (on this point, more in chapter 7).

We now have an answer to the no-luck/infallibility objection. The objection was premised on the assumption that a luck-free conviction amounted to a conviction whose correctness was infallible. While Thomson's account could be prone to the objection, I think that by distinguishing between endemic and eliminable luck, the objection loses its bite. We can reformulate Thomson's requirements as follows. A criminal conviction is acceptable provided (i) the probability of guilt on the evidence is sufficiently high, and (ii*) the trial proceeding should have eliminated any trace of eliminable luck (though endemic luck will inevitably persist persist). Requirement (i) is inherited from the probabilists, while requirement (ii*) is a modification of Thomson's requirement (ii). Importantly, requirement (ii*) does not amount to infallibility, because it openly acknowledges the inevitable role of endemic luck.

But now the question arises of how we can *eliminate* eliminable luck. Given the findings of this chapter, the answer is simple. We can eliminate eliminable luck so long as we reduce the errors of the trial system which are reducible. So, the elimination of eliminable luck is nothing bad the reduction of the trial system's errors within the limits of the system itself. And making sense of error reduction (and of the related notion of discriminating power) has been the central topic of this chapter.

CHAPTER 3

A DEBATE THAT BEGAN FORTY YEARS AGO

This chapter reviews the literature on statistical evidence and probability in the courtroom. The overview is organized historically, beginning with the sixties up to more contemporary literature. Although the dissertation concentrates on criminal trials, in this overview I shall also discuss civil trials. Historically, the academic debate has intermixed observations about civil and criminal cases, and thus for the purpose of this chapter it is best to keep the two together.

3.1 THE SIXTIES AND THE SEVENTIES

Beginning with the early sixties, a number of scholars suggested that probability and statistical methods would improve our theoretical understanding of the law of evidence and procedure. For example, one suggestion was that standards of proof be interpreted in terms of probability thresholds (Ball, 1961; Kaplan, 1968; Simon and Mahan, 1971; Lempert, 1977; Kaye, 1978). Other scholars thought that probability and statistical methods would help fact-finders weigh the evidence and reach more accurate verdicts

(Cullison, 1969; Finkelstein and Fairley, 1970, 1971). On the other hand, many remarked that standards of proof should not be probabilistically quantified, and that probability and statistical methods would impoverish and misconstrue legal reasoning, rather than improve it (Tribe, 1971a,b; Underwood, 1977; Cohen, 1977, 1981a; Nesson, 1979). In this section, I outline the first steps in this debate.

3.1.1 Kaplan and Finkelstein & Fairley

David Kaplan (1968) argued that expected utility theory could provide theoretical insights into the law of evidence.[1] Some of these insights included a justification of the appropriate threshold probability that is required for a conviction in civil and criminal cases. Kaplan's suggestion was that, in a civil case, the probability of guilt must be greater than 0.5, and in a criminal case, it must be 0.9 or higher. While these numerical values may not come as much of a surprise, it is remarkable that they were mathematically derived by applying the general principles of expected utility theory. Roughly speaking, the idea is that in civil trials convicting an innocent party is as costly as acquitting a guilty party, and thus, the probability threshold is slightly above 0.5. In criminal trials, instead, acquitting an innocent is viewed as more costly, and therefore

[1] Expected utility theory is concerned with providing rational principles for making decisions under risk and it prescribes that *rational choices are those that maximize the agent's expected utilities.* Here is an example. Mark is undecided on whether to study or not for an exam (in short, S or non-S). He is undecided whether he should study or not, because he is uncertain whether he will pass the exam or not (in short, P or non-P). So, there are in total four possible outcomes: S and P; S and non-P; non-S and P; non-S and non-P. Let U be a utility function and let P be a probability function: the former describes the utility associated with each outcome, according to some utility metric, and the latter describes the probability of each outcome, according to some interpretation of probability. Expected utility theory says that Mark should study only if the expected utility of studying is greater than the expected utility of non-studying, i.e. only if $EU(S) > EU(\textit{non-S})$, where
$EU(S) = P(P|S)U(S \textit{ and } P) + P(\textit{non-P}|S)U(S \textit{ and non-}P)$ and
$EU(\textit{non-S}) = P(P|\textit{non-S})U(\textit{non-S and } P) + P(\textit{non-P}|\textit{non-S})U(\textit{non-S and non-}P)$.

In the sixties, we witness a general tendency to apply expected utility theory, probability, and economics to the analysis of law. For an economic analysis of tort law, see Calabresi (1961); and for an economic analysis of punishment, see Becker (1968). The classical textbook on economic analysis of law is Posner (1973).

the probability threshold is higher.[2]

Kalplan assumed that an estimate of the probability of guilt could be calculated in some way, yet he did not explain in any detail how this could be done. Finkelstein and Fairley (1970) offerred an elegant answer to this problem by using Bayes' theorem.[3] In general, their purpose was to show how probabilistic tools could help the fact-finders weigh and assess the evidence and reach more accurate verdicts. As a case study, they focused on a limited class of criminal cases, those involving what we may call *statistical trace evidence*. Bloodstains recovered at the crime scene are an example of trace evidence. If the suspect or the defendant is found to have a blood type that matches the traces, this would be incriminating evidence against him, provided the blood type does not appear too frequently. That is why a statistical estimate of the blood type's frequency in a population of potential suspects is needed, whence the name 'statistical trace evidence.'

To ascertain whether the defendant is the source of the traces, one could ask whether the traces are unique or not. The problem with this approach—as Finkelstein and Fairley pointed out—is that uniqueness can hardly be proven. But even though crime traces can

[2] In particular, Kaplan begins with an assessment of the relative disutilities associated with convicting an innocent, D_i, and with acquitting a guilty defendant, D_g. Now, suppose that the probability of guilt and innocence equals P_g and P_i such that $P_g = 1 - P_i$. To convict—Kaplan suggests—the jury must be believe that

$$P_g D_g > (1 - P_g) D_i.$$

The inequality represents a situation in which the expected disutility resulting from acquitting a guilty defendant is larger than the disutility resulting from convicting an innocent defendant. So, given the inevitable possibility of error, such a situation would be one in which convicting is less harmful than acquitting, so that conviction is justified. But the inequality holds only if P_g reaches a certain value. From the above inequality, by algebra, we have

$$\frac{P_g}{1-P_g} > \frac{D_i}{D_g}.$$

This formula gives a precise indication of how high the probability of guilt must be to justify a guilty verdict, relative to the ratio between D_i and D_g. If we consider that the disutility of convicting an innocent is as harmful as the disutility of acquitting an innocent, i.e., $D_g = D_i$—as it might be the case in a civil case—, the lower bound for P_g must be at least $\frac{1}{2}$. If, instead,we think that $\frac{D_i}{D_g} = \frac{9}{1}$—as it might be more appropriate in a criminal case—, the lower bound for P_g must be at least 0.9.

[3] Kaplan just touches upon this problem; see (Kaplan, 1968, p. 1084), where Bayes' theorem and its application are briefly discussed.

rarely lead to a unique identification, they can still be very useful so long as they appear infrequently.[4] Consequently, Finkelstein and Fairley suggested to set aside discussions of uniqueness and apply Bayes' theorem as a way to assess the weight of the trace evidence in cases in which no uniqueness claim can be made. The theorem yields an estimate of a defendant's probability of guilt after taking into consideration statistical trace evidence. For the theorem to get off the ground, however, an estimate of the defendant's prior probability of guilt is also needed.

The prior probability of guilt does not mean the probability of guilt before any incriminating evidence is considered. It means the probability of guilt given the incriminating evidence which is non-statistical and non-quantitive, that is, given all incriminating evidence except the statistical trace evidence. But since prior probabilities are difficult to assess, Finkelstein and Fairley proposed to consider an *interval of possible values*. Here the idea is that if the posterior probability remains high, regardless of the assigned prior probability within the interval, this would show that the trace evidence in question is very probative. Otherwise, if the posterior probability varies significantly depending on the choice of the prior probability, this would be a sign that the statistical trace evidence is of little probative value.

(At this point a clarification is in order. As Finkelstein and Fairley remarked in a later article, they did not intend Bayes' theorem to yield a final *probability of guilt* (Finkelstein and Fairley, 1971). Bayes' theorem is only supposed to yield a probability that the defendant is the *source* of the evidence trace on the crime scene. The only question trace evidence can address, after all, is whether the defendant is the source of the traces. This clarification concerning the scope of the application of Bayes' theorem

[4] The authors wrote:

> Few, if any, evidentiary traces can be demonstrated by statistical analyses to be unique to a defendant. There is, however, a class of traces, potentially useful as evidence, which could be shown to appear only infrequently. What is the probative value of such non-unique traces? We propose to show that non-unique traces generally deserve substantial evidentiary weight, and that by the explicit use of mathematical theory the data can be cast in a form permitting more effective use of this evidence by the jury. (Finkelstein and Fairley, 1970, p. 496).

is instructive, because it highlights two different positions one can take. On one hand, one may expect Bayes' theorem to yield a final probability of guilt—and this would be a quite demanding expectation. But one may also, and more modestly, use Bayes' theorem as a formal tool for assessing intermediate questions such as whether the defendant is the source of the crime traces. Finkelstein and Fairley seemed to prefer the more cautious and more modest option, although their language was sometimes ambiguous.)

3.1.2 Tribe

The works of Kaplan and Finkelstein & Fairley are an example of the recognition which statistical and Bayesian methods received among legal scholars in the sixties and seventies.[5] But many disagreed, and Lawrence Tribe (1971b) was one of the most vocal critics. He opposed Finkelstein and Fairley's suggestion that Bayes' theorem could be used as a tool to estimate the probability of guilt. He also objected to Kaplan's idea that we could quantify standards of proof, on the ground that a quantification of guilt is morally unacceptable. Tribe's critique of Finkelstein and Fairley suggests that certain subtleties of legal reasoning are likely to be overlooked if one tries to fit everything in one schematic formula, such as Bayes' theorem. I leave the details of this critique to an footnote,[6] but the message is straightforward: there are practical and psychological

[5] See also Cullison (1969).

[6] In applying Bayes' theorem, Finkelstein and Fairley simply assumed that $P(\textit{match}|\textit{accused is guilty}) = 1$. Now, this is often the case, for if one is the criminal, he probably left a trace at the crime scene. Yet the criminal could have been very careful not to live any trace, or even if he did, the trace could have turned out not to match. This means that $P(\textit{match}|\textit{accused is guilty}) < 1$. This eventuality—Tribe argues—was overlooked because Finkelstein and Fairley were *blinded by the formalism*. They assumed that $P(\textit{match}|\textit{accused is innocent}) = f$, where f is the frequency of the matching feature in the suspect population. But this assumption overlooks the possibility that the trace evidence could have been left by the defendant without him being the actual perpetrator, whence $P(\textit{match}|\textit{accused is innocent}) > f$. See (Tribe, 1971b, p. 1361).

Further, Tribe notes that Finkelstein and Fairley talk about the probability of guilt, while the trace evidence can only have bearing on the probability that the defendant is the source of the traces. Now, *source probability* and *probability of guilt* are very different. Tribe contends that Finkelstein and Fairley confused two different questions because the Bayesian formalism forced them to rigidly conceptualize a legal question into the abstractions of a mathematical formula, thereby causing them to lose sight of the legal complexities of the case. See (Tribe, 1971b, p. 1365). To be sure, these objections could be overcome by a more careful formal Bayesian analysis; for instance, see chapter 2. In defense of Finkelstein and limits in using Bayes' theorem because we might be confused by the formalism and be induced to oversimplify the inferences toward guilt.

Tribe's second, moral critique against Kaplan is more radical. Let us suppose—Tribe concedes—that we can accurately estimate the probability of guilt. Kaplan defines an explicit threshold probability that is required to convict, e.g. 0.95 in a criminal case, leaving a 0.05 margin of explicitly measurable uncertainty. Thus, any juror who convicts would accept that for every 20 convicted individuals, one is innocent. This posture—Tribes admits—is not in itself immoral; after all, the legal system is bound to commit mistakes and we cannot lie to ourselves. The *goal* of the trial system, however, is not to convict with a 0.95 probability, but to reach the *greatest possible certainty that is reasonably attainable*. True, mistakes will occur anyway, but a system that explicitly aims at making no more than a 0.5 rate of mistaken convictions is much different from a system that, while aiming at making no mistake whatsoever, ends up having, say, a rate of 0.5 mistaken convictions. Tribe tolerates the latter and not the former system, which

Fairley, it bears saying that they were concerned with a simplified and hypothetical scenario, not with the complexities of an actual case.

Another objection involves the estimate of the prior probability of guilt. Finkelstein and Fairley suggested that the prior probability should be the result of considering the non-statistical and non-quantitative evidence. Yet here is precisely where the problem begins: how can one accurately translate the probative force of non-statistical and non-quantitive evidence into a numerical value of the prior probability of guilty? This probability is hard to assess and assigning it a value is likely to be a matter of mere guessing. In all fairness, Finkelstein and Fairley suggested that we use a range of possible prior probability values whenever a unique value cannot be found. But this suggestion does not eliminate the problem; it simply pushes it elsewhere. Within the said range there will be a lowest value, so the question would become one of how to translate the non-statistical evidence into a prior probability of guilt *which is the lowest possible*. Again, this will be very much a matter of guesswork. In the limit case in which there is no other incriminating evidence except the quantitative and statistical evidence, the prior probability—it could be suggested—should equal zero. But because of the mathematical formulation of Bayes' theorem, if the prior probability equals zero, the posterior probability will remain zero no matter what. Setting the prior probability to zero, albeit intuitively appealing, makes Bayes' theorem useless. The only alternative is to the set the prior probability to a value that is low enough. But what is 'low enough'? The question becomes more pressing if we pair it with the requirement of the presumption of innocence; this is a constitutional guarantee for any criminal defendant. A natural way to interpret it probabilistically would be to set the prior probability of guilt to zero, but we have seen that this is not possible. The prior probability of guilt, then, must be set to a small value greater than zero. But it is unclear what this value should be, and saying "a low enough value" seems to undermine the very need for using precise numerical and probabilistic methods to aid legal decision-making.

is the one resulting from adopting the Bayesian recipe. The difference can also be put in terms of *intended mistaken convictions*: in one system the mistakes are intended, while in the other they are not, although the actual rate of mistaken convictions might be the same across the two systems.[7]

In addition, Tribe notes that the trial process is *not only* an objective search for *historical truth*. It is also a *ritual*, which is embedded in societal customs and beliefs.[8] A very integral part of this ritual is the jury, performing the highly humanizing function of mediating between the law in abstract and the human needs of the defendant.[9] And the jury's task—Tribe believes—will be rendered more difficult and impractical by the numbers yielded by a mathematical formula: jurors will be intimidated by the seeming precision and inexorability of numbers, and they will be oblivious of their humanizing function.[10]

3.1.3 L. J. Cohen

If Tribe's critique of statistical and probabilistic methods in the courtroom was principally a moral one, other authors formulated critiques which we may call epistemic.[11] L. J. Cohen (1977) formulated a number of paradoxes to show that standard probabil-

[7] 'Tolerating a system in which perhaps one innocent man in a hundred is erroneously convicted despite each jury's attempt to make *as few mistakes as possible* is in this respect vastly different from instructing a jury to *aim* at a 1 percent rate . . . of mistaken convictions.' See Tribe (1971b), footnote 143, p. 1347.

[8] 'It would be a terrible mistake to forget that a typical lawsuit, either civil or criminal, is *only in part an objective search for historical truth*. It is also, and no less importantly, *a ritual*.' Emphasis mine. See Tribe (1971b), p. 1376

[9] 'One element of that ritual is the jury, an institution calculated *to mediate between the law in the abstract and the human needs*.' Emphasis mine. See Tribe (1971b), p. 1376.

[10] 'Guided and perhaps *intimidated by the seeming inexorability of numbers*, induced by the persuasive force of formulas and the precision of decimal points to perceive themselves as performing a largely mechanical and automatic role, *few jurors ... could be relied upon to recall, let alone to perform, this humanizing function*.' Emphasis mine; see (Tribe, 1971b, p. 1376).

[11] In the seventies expected utility theory and standard probability theory came under some pressure. Daniel Kahneman and Amos Tversky showed that expected utility theory was not empirically correct, although it might still be normatively sound. See Kahnemman and Tversky (1979). An attack against expected utility theory was already in air; see Ellesberg (1961). Concomitantly, some expressed discontent with the standard Kolmogorov's axioms of probability, and they proposed alternative formalizations, such as L. J. Cohen's theory of Baconian probability. Another alternative is given by Shafer (1976).

ity theory—or as he calls it, Pascalian probability—is unable to properly capture legal reasoning.[12] One of these paradoxes is known as *Gatecrasher*, as follows:

> 1000 customers enter a football stadium, and 510 of them do not pay for their ticket. Suppose a random spectator is tried for not paying. Given the number of spectators who paid and did not pay, there is a 0.51 percent probability that the spectator under trial did not pay. If we quantify the governing civil standard of preponderance of the evidence as requiring a probability of at least 0.5, a probability of 0.51 seems enough to satisfy the said standard. Yet, our intuitions about civil liability suggest that a spectator cannot be held liable for not paying on the sole basis of the number of people who did and did not pay.

The paradox was meant to suggest that an understanding of the civil standard 'preponderance of the evidence' in Pascalian probabilistic terms is inadequate.[13] To dissolve the paradox, Cohen urged a rethinking of the notion of probability. The heart of his solution rested on the notion of inductive or Baconian probability, as opposed to standard or Pascalian probability.[14] To understand inductive probability, we need to understand a related notion, *inductive support*, to which I now turn.

Inductive support applies to generalizations concerning *class* of individuals or phenomena. As Cohen puts it, the '[inductive] support for a generalization . . . is graded . . . by its degree of resistance to falsification by relevant variables' (p. 166). For every generalization, a number of relevant variables, if manipulated, could falsify it. For instance, the generalization '*whenever a dark cloud is present, rain is imminent*' could be falsified by manipulating variables such as temperature, pressure, geographical position, time of day, season. The degree of inductive support of a generalization is proportional

[12] In Cohen (1977), six difficulties are mentioned; see part II, chapter 4–10. These difficulties concern: conjunction; inference upon inference; negation; proof beyond reasonable doubt; criterion; corroboration and convergence.

[13] The paradox was also meant to put pressure on the negation rule that $P(\neg A) = 1 - P(A)$. Cohen thought that this rule creates evidence from the absence of evidence. The idea is this: If I have little evidence for A, its probability will be low; by the negation rule, the probability of the negation will be extremely high.

[14] The adjective ' Baconian' is in honour of the philosopher Francis Bacon (1561-1626) who gave an account of the inductive method in terms of *eliminative induction*; see his *Novum Organum* (1620).

to the number of relevant variables that have been manipulated: the more variable manipulations the generalization resists, the more inductive support it has.[15] We are now ready to move from the inductive support of a generalization to the inductive probability of a single event: the inductive probability of an event equals the inductive support of the relevant generalization. For instance, if we see dark clouds in the sky and we wish to know the inductive probability of the event '*rain is imminent*,' we can invoke the generalization '*whenever a dark cloud is present, rain is imminent*.' If the generalization has degree d of inductive support, the event in question will have degree d of inductive probability.[16]

Let us now see Cohen's solution to the Gatecrasher paradox. While the standard probability of '*the spectator did not pay for his ticket*' equals 0.51, Cohen argued that the inductive probability is close to zero. His argument is roughly as follows. The relevant generalization is something like '*if one is a spectator, he does not pay for his ticket*.' Now, such a generalization—Cohen thought—has little inductive support. The generalization has some inductive support, but only based on the accidental circumstance that

[15] If after manipulating one relevant variable, the generalization has withstood falsification, it has minimal inductive support; if after manipulating all relevant variables, the generalization has again withstood falsification, then it has full inductive support. Numerically and more generally, let n be the number of all relevant variables for a generalization, if the generalization has survived falsification against the manipulation of m variables, its inductive support equals m/n; if it has survived the manipulation of *all* relevant variables, its inductive support equals one. This procedure is called, not surprisingly, *method of the relevant variables*. This method has many historical predecessor, from Francis Bacon method of eliminative induction to J. S. Mill method of agreement and difference. And more generally, this seems to be—roughly speaking—the method that statisticians adopt whenever they want to establish non-spurious causal connections between events, e.g. smoking and cancer.

[16] More generally, suppose we have available a generalization of the form $\forall x(\varphi(x) \rightarrow \psi(x))$, with a degree d of inductive support. Suppose, also, $\varphi(a)$ obtains, with a some individual. The question is now whether $\psi(a)$ obtains, as well. Cohen's answer is that $\psi(a)$ has a degree d of *inductive probability*, because the applicable generalization has a degree d of *inductive support*. Notice that inductive probability does not follow the Kolmogorov's axioms of probability. Take, for instance, the law of negation, saying that $P(X) = 1 - P(\neg X)$, with P a classical probability function. Suppose that the generalization $\forall x(Y(x) \rightarrow X(x))$ has a degree d of inductive support, and also that $Y(a)$ obtains. Then, $X(a)$ has a degree d of inductive probability. It does not follow, however, that $\neg X(a)$ has a degree $1 - d$ of inductive probability. This would be the case only if another generalization of the form $\forall x(Y(x) \rightarrow \neg X(x))$ were available, with a degree $1 - d$ of inductive support. But such a generalization might not be available, and even when available its degree of inductive support need not equal $1 - d$.

510 out of 1000 people did not pay for their ticket. Yet, had the people at the stadium been different, or had the stadium been in a different place, etc., the ratio between paying and non-paying spectators might have changed.[17] The generalization in question, then, possesses little resistance to falsification against relevant variables, and thus it has little inductive support. It follows that the conclusion that the spectator did not pay has a low degree of inductive probability, although it appears to have a higher degree of standard probability. In the end, if standards of proof are understood in terms of inductive probability, the above reasoning should be enough to block a finding of guilt.

3.1.4 Nesson

After a moral and an epistemic critique of the application of probability and statistical methods to the courtroom, I turn to a critique which we may call socio-political. Charles Nesson (1979) constructed a hypothetical scenario *à la* Cohen which applies to criminal trials, as follows:

> *Prisoners*. “In an enclosed yard are twenty-five identically dressed prisoners and a prison guard. The sole witness is too far away to distinguish individual features. He sees the guard, recognizable by his uniform, trip and fall, apparently knocking himself out. The prisoners huddle and argue. One breaks away from the others and goes to a shed in the corner of the yard to hide. The other twenty- four set upon the fallen guard and kill him. After the killing, the hidden prisoner emerges from the shed and mixes with the other prisoners. When the authorities later enter the yard, they find the dead guard and the twenty-five prisoners. Given these facts, twenty-four of the twenty-five are guilty of murder.
>
> Suppose that a murder indictment is brought against one of the prisoners—call him Prisoner 1. If the only evidence at trial is the testimony of our distant witness, it would seem that a verdict of acquittal must be directed for the defendant. The prosecution’s best case is purely statistical. Nothing distinguishes Prisoner i from the other twenty-four prisoners. The odds may be twenty-four in twenty-five that the defendant was one of the murderers,

[17] Cohen is not so clear on this point in his 1977 book. He is more explicit in a later article; see (Cohen, 1981a, p. 634).

> but there is no way, on this evidence, that a jury could form an "abiding conviction" that the defendant was guilty. A conclusion that Prisoner 1 was guilty, a court would say, could be based only on speculation, for there is no basis in the evidence for differentiating the defendant from the other prisoners."[18]

With a few variations, this is essentially the scenario *Prisoners* we saw in chapter 1. The scenario was meant to suggest that the criminal standard of proof 'beyond a reasonable doubt' should not be understood in terms of a threshold probability of guilt.[19] If the prosecutor's case against Prisoner 1 were to rest solely on the statistics that 24 prisoners participated in the killing, Nesson argued that the prosecutor's case should be dismissed and should not even go to a jury, because no jury could possibly convict Prisoner 1. This is puzzling, Nesson adds, because the guilt probability of Prisoner 1 is extremely high. In contrast, suppose the prosecutor's case is based on two items of evidence: first, relatively weak testimonial or circumstantial evidence against the defendant; second, statistical evidence that the defendant's probability of guilt equals 0.5 (because there were only two prisoners and only one killed the guard). We can concur that the probability of guit in this new scenario is higher than 0.5 but not as high as in the original scenario. Now, the "true" puzzle is that, although the defendant's probability of guilt in the original scenario is higher, the prosecutor's case is allowed to go to the jury only in the second scenario. The reason is that a jury is allowed to assess the weight of circumstantial or testimonial evidence. So, a defendant could potentially be convicted, despite the fact that the evidence against him is probabilistically weaker. Or at least this is what Nesson holds.[20]

[18] (Nesson, 1979, p. 1192).

[19] 'The thesis pursued here is that any structure which reduces criminal cases to a simplified assessment of what might be called the "chances of guilt" is fundamentally at odds with the concept of reasonable doubt, and hence to be dismissed as a mode of determining the ultimate question of guilt.' See (Nesson, 1979, p. 1192).

[20] 'Why should it be that the high likelihood but starkly numerical case is thrown out of court while cases based on self-serving testimony or additional circumstantial evidence will be put to the jury? The question becomes truly puzzling when one considers that even a case in which the quantifiable likelihood of guilt was much lower—for example, when originally only two prisoners were in the yard—might be allowed to go to the jury as long as the prosecutor's case was bolstered by additional circumstantial evidence.' (Nesson, 1979, p. 1194).

If Nesson is correct, an alarming question suggests itself: Why are legal proceedings so indifferent towards the probability of guilt? Nesson's response is rather cynical, and even more alarming. Contrary to popular belief, the goal of the trial system would not be to determine the truth about particular events;[21] rather, its primary goal would be to settle disputes as effectively as possible and to create a sense of "public deference" towards the trial system and its decisions.[22] Interestingly enough, if a clear-cut decision rule is adopted (e.g. convict in a criminal case if the probability of guilt is above the 0.99 probability threshold), the decisions' procedures could be scrutinized and examined more easily than if the rules were complex; and consequently, public deference towards the trial system would diminish. Lawyers and judges—Nesson suggests—prefer complexities, contradictory testimonies, and ambiguities in the evidence, because these effectively shield them from criticism and are more congenial to the goal of public deference. If this is the hidden goal of the trial's system, Nesson's scenario is too simple a case, and letting a jury decide it would make the jury too easily susceptible to criticism and public scrutiny.[23]

3.2 Eighties

In the eighties the interest in questions concerning the relationship between legal evidence and probabilistic methods grew dramatically (Kaye, 1979a,b, 1980, 1982, 1986a,b,c, 1989; Saks and Kidd, 1980; Gärdenförs et al., 1983; Nesson, 1985; Allen, 1986; Lempert, 1986; Schmalbeck, 1986; Thomson, 1986; Zuckerman, 1986; Friedman, 1987;

[21] That this was the goal of the trial system had been already challenged by Tribe in his 1971 article.

[22] 'Many of the procedures of our legal system are best understood as ways to promote public acceptance of verdicts.' See (Nesson, 1985, p. 1368).

[23] In the title of his 1979 article we find the expression '*the value of complexity*,' which very well summarizes Nesson's point. In a later 1985 paper, Nesson developed his argument by suggesting that trial decisions gain acceptance from the public only if they are perceived as having to do with the events rather than with the evidence. In other words, in order for a trial decision to have authority, it must assert itself as a decision about the truth of the events disputed during the trial proceedings, and not as a decision about how much the evidence proved the events. According to Nesson, decisions based on an explicitly quantified standard of proof are perceived by the public as being about the evidence, and not the event; and therefore they would gain little authority or acceptability. See Nesson (1985).

Schoeman, 1987; Thompson and Shumann, 1987; Dant, 1988; Wright, 1988; Shaviro, 1989). In 1986, a very important evidence law conference took place and its proceedings were later published in a volume under the title *Probability and Inference in the Law of Evidence* (Tillers and Green, 1988). Many of the authors whose arguments and ideas I am going to review participated in that conference.

3.2.1 Thomson

The moral philosopher Judith Thomson (1986) considered a number of hypothetical scenarios similar to Cohen's Gatecrasher and Nesson's Prisoners. It is useful to recount one of them, known as *Blue Bus hypothetical*:

> *Blue Bus*. In a hit-and-run accident, a color-blind person is run down by a bus in a small town. There are no other witnesses. In the small town, Blue Company operates 80 percent of the buses, while Red Company operates the remaining 20 percent. The victim sues Blue Company and argues that it is liable with a 80 percent chance, because it operates 80 percent of the buses. Given that the standard of proof in a civil case is preponderance of the evidence, which can be quantified as requiring to establish liability by at least 50 percent, Blue Company should be convicted. Yet, our intuitions about civil liability seem to suggest otherwise: Blue Company cannot be held liable because it operates more buses than Red Company.

The problem raised by this scenario is that, although the probability that Blue Company is responsible for the injury is quite high, and it exceeds the 0.5 threshold, it seems unfair to convict Blue Company. Thomson argued that the evidence against Blue Company is epistemically deficient because it is not *individualized evidence*. Her account of individualized evidence comes down to the idea that a piece of evidence is individualized if it is causally connected with the event disputed at trial.[24] The evidence against Blue

[24] Here Thomson seems to invoke the notion of causality to give an account of the individualized or case-specific nature of a piece of evidence. This move can be contested, however. If a causal relation is interpreted as holding generally, or among a class of phenomena, it need not hold between the specific evidence proffered at trial and the specific event under dispute. The existence of a causal relation between the evidence and the event to be proven does not entail that this relation hold in the *individual case* of interest for the trial; causality does not seem to be a guarantee of the individualized nature of the evidence.

Company consists only in how many buses it operates. This is causally irrelevant for the occurrence of the disputed event. Hence, the evidence in question is non-individualized.

But why is non-individualized evidence not enough for finding liability? Thomson drew a parallelism between knowledge claims and convictions: for a subject to (claim to) know a proposition p, she must at least have a good enough reason to believe p; similarly, for a jury to convict, it must have a good enough reason to believe that the defendant is guilty. Next, Thomson suggested that a *good enough reason* should be one that *guarantees* the truth of p, so that it would *not* just be a matter of *epistemic luck or accident* whether a subject is right about p.

In the Blue Bus scenario the only reason for finding liability is that Blue Company owns more buses than Red Company and that it is is more likely to be liable. Thomson reasoned that the high chance that Blue Company is liable is not a guarantee that the company is in fact liable because, on the sole ground of how many buses it operates, it would be a matter of accident whether Blue Company is liable.[25] As a guarantee for the truth of a conclusion, we've seen, Thomson suggested that a *causal connection* should hold between the evidence and the factual conclusion to be inferred from the evidence.[26] Thus, absent a causal connection between the companies' ownership distribution and the injury suffered by the victim, the jury was not given evidence guaranteeing that Blue Company was liable.[27] On Thomson's account, a piece of evidence is a good enough

To obviate to this difficulty, it seems plausible to require that the causal relation be an *individualized causal relation*, namely one occurring in the specific circumstances under dispute. On individualized causation, see the manipulability account in Woodward (2003).

[25] Thomson gave the example of a fair lottery. With a sufficiently large number of tickets, the chance that a certain ticket, say ticket number 3456678, loses is extremely high, yet it would be a matter of accident and (bad) luck if ticket 3456678, in fact, loses. There are cases such as the lottery case in which drawing factual conclusions on the basis of the high probability alone and being right about them is only a matter of accident or (good or bad) luck.

[26] Thomson did not offer a detailed account of the notion of causality. She merely relied on intuitive and under-specified notion of cause. But notice that in linking evidence and causation, Thomson is not alone. In dealing with lottery propositions and other problematic cases, Achinstein (1978) argues that a causal explanation between the evidence and the hypothesis is required.

[27] As Thomson put it,

> [I]f a jury declares a defendant guilty on the ground of non-individualized evidence alone, then it is just luck for the jury if what it declares true is true—and . . . it [would be] unjust to impose liability when that is the case. . . . What would make it not to be just luck for the jury

reason to draw a factual conclusion at trial whenever (i) it guarantees the truth of the conclusion, i.e. it is causally connected with the events described by the conclusion, and (ii) the jury believes that the casual connection holds with a sufficiently high degree of probability, according to a probabilistic interpretations of standards of proof.

(Thomson cautioned that she did not offer a definition of individualized evidence consisting of necessary and sufficient conditions. She pointed to a type of evidence that despite failing to be causally connected with the disputed event can still count as individualized. Suppose that a murderer is known to be one-legged, left-handed, entirely bald, and extremely tall, and that the a biologist-statistician testifies that people having such features exist with a frequency of 1 in 10 million.[28] If the defendant possesses those rare features, the jury may not conclude that the defendant is the murderer because, taking a population of 30 million people as the range of possible suspects, we would expect three people having those features. However, if the frequency of the identifying features in question were much lower, or if the statistician could testify as to their uniqueness,[29] and the defendant possessed them, this would suggest that the statistical evidence singles out one individual and it is therefore individualized.[30] Hence, evidence that is not causally connected may still be individualized provided it uniquely identifies the perpetrator.)

3.2.2 Zuckerman.

Adrian Zuckerman (1986) focused on a question that went unnoticed so far in the debate: the distinction between matters of fact and matters of law. The fact-law distinction—he observed—is taken for granted in the debate about probability and the law; the probabilities are taken to apply to brute facts and to exist independently of the law. For

> if what it declares true is true? A guarantee. I suggested that individualized evidence for a defendant's guilt is evidence which is in an appropriate way causally connected with the (putative) fact that the defendant is guilty, and hence (putatively) guarantees the defendant's guilt. (Thomson, 1986, p. 214)

[28] See (Thomson, 1986, p. 215).

[29] Thomson imagines a far-fetched situation in which the defendant is known to have only one eye.

[30] This is true if we disregard the doubts as to whether the statistician's claim that the features in question are unique holds good.

example, in the Blue Bus scenario, the tacit assumption is that there is a 0.8 probability that a bus from Blue Company injured the victim. The fact *bus-from-Blue-Company-injures-victim* might seem to exist independently of the law, yet closer scrutiny suggests otherwise. The fact '*injury*' does not exist independently of the law, for an action and its consequences constitute an injury only so long as the law counts the action and its consequences as such. As Zuckerman puts it, 'the ultimate facts are only because the law draws our attention to them by attaching legal consequences to their existence or non-existence' (p. 492). Zuckerman's point is that which facts should be proven are the facts that the law determines should be proven. Further, not only the law selects *which* facts are relevant and should be proven; law and morality also constrain *how* facts should be proven.

Next, Zuckerman observed that the fact-finders are expected to *adjudicate on the merits* by taking into consideration 'all the infinitely varying circumstances of the individual litigants' (p. .496) Now, finding Blue Company or the spectator in Gatecrasher liable would seem to contravene to the principle of adjudication on the merits, so that no verdict should be issued in these scenarios. For instance, about Gatecrasher, Zuckerman argued that a finding of guilt would amount to endorsing the principle of *corporate liability* according to which one is found guilty depending on what others have done and not him; and this patently contradicts adjudicating on the merits. Importantly, Zuckerman is justified in claiming that corporate liability should not be used to establish guilt because he previously argued that the way in which ultimate facts are proven should be informed by moral and legal principles, such as the principle of adjudication on the merits.

3.2.3 Kaye

I've reviewed the epistemic critique by Thomson based on causality and the moral critique by Zuckerman. These critiques elaborate on the three critical strands (epistemic, moral, and socio-political) that we've already encountered in the seventies (Cohen, Tribe, Nesson). I now turn to two authors who favour probability and statistical methods in the courtroom. The first is David Kaye, one of the most distinguished scholars of statistics and the law. Kaye (1979a,b, 1981, 1982, 1986b) challenged the assumption that, in the Gatecrasher scenario, the probability that the spectator on trial did not pay for his ticket is 0.51. The value 0.51 seems to follow from the fact that 510 spectators did not pay for their tickets out of a total of 1000 spectators. Kaye endorsed a Bayesian subjective view of probability and he contended that the probability should be lower than 0.51. The reason is that the plaintiff in Gatecrasher did not introduce any evidence other than the naked statistical frequency of paying and non-paying customers. The plaintiff's failure to adduce any more particularized evidence should lower the final probability of guilt and bring it below the critical threshold of 0.5.[31]

(A note on terminology is in order. The expressions '*naked statistical evidence*' and '*particularized*' or '*individualized evidence*' are common among legal scholars. The former expression typically designates the statistical evidence available in Gatecrasher, Prisoners, and Blue Bus. In general, it designates statistical or quantitative evidence which is unaccompanied by other non-statistical evidence, whence the adjective 'naked.' The expression 'individualized' or 'particularized' refer to evidence, such as eyewitness evidence or trace evidence, which is connected to the specific fact under dispute. These are vague characterizations and some scholars contest their intelligibility, but I hope they are clear enough for the present discussion.[32])

On Kaye's proposal, the probability that the spectator did not pay for his ticket does not exceed the threshold of 0.5. But why? It is general legal practise that a party's failure to adduce evidence *which the party is reasonably expected to introduce* should weaken the party's case. This means that a plaintiff is penalized on the assumption that his failure to introduce more evidence was an attempt to conceal evidence that would otherwise favour the defendant. In probabilistic terms, penalizing the plaintiff amounts to lowering the probability of guilt.[33] Such lowering of the probability of guilt

[31] This idea is first expressed in (Kaye, 1979a, p.).

[32] For instance Saks and Kidd (1980) deny that there is any epistemic distinction between naked statistical evidence or base-rate information and particularized or case-specific evidence.

[33] For a probabilistic analyses of the problem of missing evidence, see Kaye (1986b). In civil cases, the doctrine of *res ipsa loquitur* is another example of how one party's failure to introduce evidence that it was not negligent may result in a finding of liability; in other words, under certain circumstances, negligence is presumed unless proven otherwise. This doctrine is supposed to encourage the party that has access to

is a *matter of policy*, aimed at creating an incentive for plaintiffs to produce the most complete available evidence, so that fewer mistakes are committed in the long run.[34] In Gatecrasher, Kaye's suggestion that the failure to introduce more particularized evidence should lower the probability of guilt is simply an instance of such general practise.[35]

However, Kaye admits, the plaintiff might not be able to introduce more particularized evidence, because no other evidence is in fact available, and if so he should be penalized. To address this complication, Kaye distinguishes two classes of cases: those in which naked statistical evidence is *justifiably* the only incriminating evidence, and those in which naked statistical evidence is *non-justifiably* the only incriminating evidence. In the latter, it does not seem unfair to penalize the defendant and lower the probability of guilt on the evidence.[36] The probability of guilt on naked statistical evidence, instead, should not be lowered, whenever naked statistical evidence is—*justifiably*—the only incriminating evidence. In this sort of cases, Kaye argued that a probability of guilt greater than 0.5 should be enough for finding liability in civil cases.

Finally, Kaye added an important dimension to the debate about probability and the law. He emphasized that, in taking a stance in the debate, we should make explicit the values and the goals of the trial system. For instance, Kaye formulated a general argument defending the *maximum likelihood rule*, i.e. the rule that a conviction is warranted

the disproving evidence to introduce such evidence. For a discussion of *res ipsa loquitur* in relation to probability, see Jaffe (1952) and Kaye (1978); for a discussion of the doctrine in relation to the problem of presumptions, see Allen (1994).

[34] Kaye thinks that if more evidence is considered at trial, decisions will be more informed and thus less likely to be erroneous.

[35] Kaye suggested that the policy be adopted that 'where individualized evidence is likely to be available—evidence which would typically permit better estimates of the probabilities than can be had by background statistics alone—plaintiff should be forced to produce it. In the long run, fewer mistaken verdicts should result under this rule of law; see (Kaye, 1979b, p. 40). Along the same lines, see (Kaye, 1980, p. 610).

[36] '[T]he legal issue created by naked statistical evidence can be resolved satisfactorily if it is granted that in most cases probative, non-quantitative evidence should also be readily available. In these circumstances, a rule requiring that at least some such evidence be brought to bear on the case is not so mysterious. It is merely a device calculated to enhance the accuracy of the fact-finding process in a manner that is fair for both parties and that is not overly burdensome to the proponent of the statistical evidence.' See (Kaye, 1982, p. 488).

when the proven probability of guilt exceeds the threshold probability of 0.5.[37] Kaye's argument was that the maximum likelihood rule is the decision rule which minimizes expected losses compared to all other possible decision rules.[38] With this in mind, the open question is whether or not the goal of the trial system is to minimize expected losses. The maximun likelihood principle might be adequate for civil but not criminal trials, in which the goal might be to minimize the defendant's losses, not overall losses, so that a threshold probability higher than 0.5 would be preferable. Be that as it may, Kaye's point is that those who disagree with him should specify which values civil or criminal trials should pursue and why these values would discourage or forbid the application of probabilistic and statistical methods in the courtroom.

3.2.4 Lempert

Richard Lempert (1986) contested that Gatecrasher, Blus Bus, or Prisoners posed a problem for a probabilistic interpretation of the standards of proof.[39] For instance, concerning Gatecrasher, he contented that Cohen's intuitions that the plaintiff should not recover damages were incorrect. Lempert, in fact, thought that there were no well-established intuitions because Gatecrasher-like cases were so artificial that no intuition about their correct legal outcome could have formed. He argued that '[i]f we encountered pure

[37] See Kaye (1982).

[38] Kaye compared three possible decisions rules: (R1) the no recovery rule; (R2) the rule that allows for recovery if the threshold of 0.5 is met; and (R3) the expected value rule. Rule (R1) mandates that in any court case in which plaintiff has only made his case in probabilistic terms, no matter what this probability is, the plaintiff should not recover any damages. This rule aligns with Cohen's position that in the gatecrasher hypothetical the tried spectator should not be convicted. Rule (R2) corresponds to the maximum likelihood principle, allowing plaintiff to recover damages if he can prove guilt according to at least 0.51 probability. Finally, rule (R3) always allows plaintiff to recover damages, but the amount recovered is weighted by the probability with which he could prove his case. Kaye showed that rule (R2) is the one that *minimizes expected losses*, provided monetarily equal losses are considered equally harmful for the plaintiff and the defendant. (Some times, however, if the plaintiff (a middle-class family) is economically weaker than the defendant (a huge corporation), a loss of equal monetary value would be more harmful for the plaintiff than it would be for the defendant.)

[39] Lempert is an evidence scholar who in the seventies had proposed a probabilistic and Bayesian interpretation, in terms of likelihood ratio, of the legal notion of *relevant evidence*. He gave a probabilistic interpretation of article 401 of the *Federal Rules of Evidence*. See Lempert (1977).

naked statistical inference cases like the gatecrasher hypothetical with any frequency, both our intuitions and the law's rule would change.'[40] Lempert conjectured that, if cases involving naked statistical evidence became very frequent, our intuitions and the law would converge toward a rule of proportionate recovery and liability.[41]

There are two interconnected points here. One is that Gatecrasher-like scenarios are artificial and tell us very little about legal cases; these scenarios would be worth considering only if they arose with a certain frequency. The second element is methodological. Lempert thought that for a given type of case, e.g. a Gatecrasher-type case, its iterations in the long run afford us an adeguate perspective on how the legal system should respond, otherwise considering only one instance of the case is uninteresting.

3.3 NINETIES

In the nineties scholars' interest in the role of statistics and probability in the courtroom faded away. At least, scholars lost interest in a debate centering around a few hypothetical scenarios that seemed to have little or nothing to do with actual legal cases. Further, in the nineties the criminal justice system changed radically. A new and powerful form of statistical evidence, DNA evidence, gained momentum in the courtroom, and this showed even more clearly the shortcomings of a debate focusing on hypothetical scenarios. The nineties were characterized by more careful analyses of the role of statistics and probability in the courtroom (Koehler and Shaviro, 1990; Allen, 1991; Fienberg and Straf, 1991; Feinberg and Kaye, 1991; Robertson and Vignaux, 1993; Schum and Kadane, 1996; Dekay, 1996; Kaye, 1999; Taroni and Aitken, 1997). Unsurprisingly, many articles were published on DNA evidence, its legal and statistical pros and cons (Kaye, 1993; Koehler, 1993; Koehler et al., 1995; Koehler, 1996a; Lempert, 1993; Robertson and Vignaux, 1995; Balding and Donnely, 1996). Psychologists become interested in naked statistical evidence and they tested whether people do have the intuitions we think they have when they are confronted with scenarios such as Blue Bus,

[40] See (Lempert, 1986, p. 461).

[41] Lempert has in mind the role of market-share liability formulated in the *Sindell* decision.

Gatecrasher, or Prisoners (Wells, 1992). I shall not review this literature here. For my purposes, I concentrare on another critique of probability and statistical methods in the courtroom, a critique rooted in moral considerations.

3.3.1 Wasserman

Tribe in the seventies argued that grounding a conviction on “numbers alone” is immoral because it mechanizes and dehumanizes the trial process. Tribe’s argument amounted to a wholesale rejection of probabilistic methods in the courtroom. After Tribe, more careful arguments have been offered, most notably by David Wasserman (1991). For him, the concern was no longer the quantification of standards of proof. Rather, Wasserman focused on the evidence used as a basis for a criminal conviction and argued that statistical evidence poses problems when (and only when) it is in tension with the presupposition embedded in the justice system that each defendant has freedom of choice.

In a criminal case, proving guilt consists in proving that the defendant committed the wrongdoing charged. If the prosecutor’s case were to rest entirely on evidence concerning the actions of others, or on evidence concerning the defendant’s previous wrongdoings, this method of proof would overlook the defendant’s individuality and freedom of choice. The reason is that presumably the defendant was free to commit or refrain from committing the wrongdoing charged: establishing that he committed it by relying on the behaviour of others, or on his previous actions, presupposes an implicit denial of the defendant’s freedom of choice. On Wasserman’s account, it is this denial that is morally objectionable.

On the account just outlined, the moral unacceptability is not a direct consequence of the statistical nature of the evidence *per se*. Character evidence is non-statistical, but its use to establish guilt would be morally objectionable because the evidence is about the defendant’s previous conduct, criminal or otherwise. For similar reasons, statistical evidence is morally unacceptable so long as it concerns the behaviour of people other than the defendant or the defendant’s past actions. The Prisoners scenario is a good illustration: If the prisoner on trial were convicted just because 99 out of 100 prisoners participated in the killing, his freedom of choice would be ignored.

3.4 21ST CENTURY

In the 21st century the interest in statistical and probabilistic methods in the courtroom grew significantly. In particular, two recent monographs on the foundations of evidence law devoted entire chapters on the topic.[42] Here I concentrate on the literature immediately relevant to the question of whether standards of proof can be interpreted probabilistically.

3.4.1 The reference class problem

A number of authors, e.g. Colyvan et al. (2001) and Allen and Pardo (2007), drew attention to the reference class problem. To illustrate, take John who is a smoker but does sports every day. For smokers the probability of dying of lung cancer is, say, 70 percent, so John has a 70 percent chance of dying of lung cancer. However, for people practising sport everyday the probability of dying of lung cancer is low, so John also has a low probability of dying of lung cancer. Depending on the reference class—that of smokers or of those practising sport—the probability of John's dying of lung cancer is significantly different. The reference class problem results from the fact that the same individual falls under multiple reference classes, each plausible and each yielding a different probability estimate.

In Blue Bus—but similar considerations apply to Gatecrasher and Prisoners—the reference class consists in the set of buses circulating in town. But why weren't other reference classes considered, such as the class of the buses operating at the time of the accident, or the class of buses that had been previously involved in accidents? Since there are multiple reference classes, different and perhaps conflicting probabilities can be derived from them. It is therefore questionable that there is 80 percent probability that the bus involved in the accident belonged to Blue Company.

[42] See (Stein, 2005) and (Ho, 2008).

3.4.2 Stein

So far we've seen epistemic, moral, and socio-political criticisms. Stein (2005), instead, mixes epistemic and procedural considerations. To begin with, we should understand the difference between probability and weight. Any probability statement has the form 'there is n percent probability that p,' where p is a proposition describing an event of interest. According to Stein, any probability statement has two dimensions: probability and weight. The notion of weigh is a function of the size of the evidentiary or factual basis supporting the probability statement. The wider the evidentiary basis, the weightier the probability statement.

Suppose now we are given a statement that is highly probable but has little weight. Should we believe it or suspend judgment? Stein thinks that epistemology alone cannot give an answer. Considerations of political morality, instead, can suggest that the suspension of judgment is to be preferred, but there is no intrinsic epistemic reason for why this should be so.[43] Let us now assume that that the correct policy to adopt is what Stein calls the weight criterion, which mandates that any legal finding be sustained by a statement that is both weighty and sufficiently probable. By using the weight criterion, it is now possible to explain why a conviction is unacceptable in scenarios like Blue Bus, Prisoners, or Gatecrasher. For instance, in Blue Bus the statement that there is 80 percent probability that the bus from Blue Company caused the accident has no weight. Its evidentiary basis concerns only the frequency of how many buses are operated by one or the other company and it does not relate to the accident under dispute. By applying the weight criterion, then, no finding of liability would be acceptable against Blue Company. Similar considerations apply to Prisoners and Gatecrasher.

In Stein's argument, the weight criterion is not the final word. In fact, a complication arises. Suppose an eyewitness testifies that a bus belonging to Blue Company hit the victim. The question is now whether this testimony is sufficient to sustain a finding

[43] 'There is nothing in epistemology that prescribes indecision as the best course of action for this type of cases [i.e., cases with non-weighty probability estimates]. Epistemology is an evaluative theory of knowledge, not a prescriptive theory of action...Epistemology cannot tell fact-finder that making an unevidenced [i.e., non-weighty] probability assessment is worse than being halted in indecision.' See (Stein, 2005, p. 84).

against Blue Company. The weight criterion requires the incriminating evidence to be weighty. Stein points out that the eyewitness evidence in itself is not weighty because one would need further evidence showing that the witness is reliable and truthful. In absence of such additional evidence, the eyewitness testimony would not be weighty. Consequently, the application of the weight criterion seems to block any inference based on eyewitness evidence as well as statistical evidence. This is a paralyzing result. To overcome the impasse, Stein formulates the principle of maximal individualization, which has two parts: one requires that all relevant evidence be considered, and the other requires that any finding against a litigant be based on evidence that was exposed and survived a procedure called maximal individualized examination.[44] Such procedure is nothing else than cross-examination, and it has the function of putting an end to the infinite open-ended nature of any evidentiary inference: If a piece of a evidence survives cross-examination, it can be deemed 'procedurally weighty' and therefore sufficient to support a verdict.

Now, cross-examination is a procedure that is particularly suited for eyewitness evidence. Through it, the truthfulness of the witness can be scrutinized. If the witness survives cross-examination, his statement can be enough to sustain a verdict. This holds good, even though the witness testimony is not weighty, but it becomes 'procedurally weighty' after cross-examination. In case of Blue Bus, instead, cross-examination cannot function because statistics about the number of buses in circulation cannot be properly cross-examined, given that the number of buses is fairly uncontroversial. Hence, a finding against Blue Company would be unacceptable.

3.4.3 Allen, Pardo, and Leiter

Allen and Leiter (2001) conceive of the courtroom as a confrontation between competing theories of the crime, theories offered by the defendant and by the prosecutor. The

[44] 'This principle unfolds into two specifici requirements. First, fact-finders must receive and consider all case-specific evidence pertaining to the case. Second, fact-finders must not make any finding against a litigant, unless the argument generating this finding and the evidence upon which this argument rests were exposed and survived maximal individualization examination.' See (Stein, 2005, p. 100).

theory to be selected should be the most plausible one. Plausibility is a notion which overlaps only partly with probability; the most probable theory need not be the most plausible one. In a similar vein, Pardo and Allen (2008) think that the strength of a theory is measured not on its mere probability, but on how well it explains the evidence. The theory that best explains the evidence will the most plausible and the one that should be selected. The plausibility of a theory depends on different variables: probability, detail, persuasiveness, capacity to accommodate the evidence introduced by both sides, simplicity, etc. On this account, the difference between civil and criminal trials is that in the former the theory to be selected is simply the most plausible among the available theories (Allen, 1986). In criminal trials, the theory to be selected is the most plausible and the one for which there is no other more plausible alternative (Allen, 2010).

Turning now to Gatecrasher, Prisoners, and Blue Bus, these are all cases in which the prosecutor advanced a probable theory, but a theory which is not sufficiently plausible and which does not explain the evidence. Take Blue Bus: the theory is, presumably, that a bus from Blue Company hit the victim. Does this theory explain the evidence? It seems not; the evidence is that Blue Company operates more buses than Red Company, but the theory that a bus from Blue Company hit the victim does not seem to explain the market share distribution between the two companies. Incidentally, this line of argument is roughly similar to Thomson's causality-based account.

3.4.4 Ho

Ho (2008) advances an account that is quite eclectic: it synthesizes many different ideas in a coherent theory. His core idea is that a conviction cannot rest on a merely probabilistic belief of guilt; it has to rest on a *categorical belief*. Scenarios such as Prisoners, Gatecrasher, and Blue Bus exemplify cases in which the belief of guilt is merely probabilistic and not categorical. But what is a categorical belief? In Ho's view, a story of the crime must be contrasted against other stories and their relative plausibility. Only if a story proves itself superior to the rival stories, we may arrive at a categorical belief in its truth. Further, because of the dialogical nature of legal fact-finding, Ho observes that the competing stories will be compared by the fact-finders. So, if one story is supported by mere statistical evidence, it will not persuade the fact-finders, and thus it will not be enough to support a verdict.

3.4.5 Pundik

Wasserman explains why we are (and should be) resistant to convicting on statistical evidence alone by an appeal to the defendant's freedom of choice, which would be undermined in using statistical evidence to establish guilt. This appeal is intuitive enough, yet it is unclear why undermining the defendant's freedom is morally objectionable. In developing Wasserman's argument, Amit Pundik (2009, 2011) has rightly emphasized that the trial system owes respect to the defendant's freedom *because* it engages in attributions of culpability and blameworthiness. Establishing that the defendant committed the wrongdoing charged while undermining his freedom, is not morally objectionable *per se*, but it becomes so insofar as it involves an attribution of culpability. Blaming a defendant for his wrongdoing, while neglecting his freedom of choice gives rise an inconsistency.

Pundik makes two assumptions: that criminal trials engage in attribution of culpability against defendants; and also, that there is a close connection between culpability and freedom. Both assumptions are plausible, especially in criminal law. Convictions are often followed by a punishment in the form of a temporal deprivation of the defendant's individual liberties, but punishment and incarceration would make little sense if defendants were not considered culpable. Further, criminal law recognizes a close connection between culpability and freedom because the prosecutor is expected to prove the *actus reus* and the *mens rea*: the former is the crime's external or physical component, while the latter is the crime's internal or mental component, the "guilty mind." No conviction can be issued in absence of the *mens rea*: if one accidentally kills a victim, with no guilty mind, the trial system will not convict him. This is generally true, although there are many complication and culpability is not a simple concept. Complications aside, for Pundik, the acceptability of a conviction depends on our account of culpability, so for him the problem of statistical evidence and probability in the courtroom is part of the larger problem of what it means to hold someone culpable.

3.5 WHAT NEXT?

In the literature I've just reviewed, we can identify different lines of argument. Some authors (Cohen, Thomson, Stein) note that statistical evidence is epistemically deficient, because it fails to be causally connected or it is not weighty. Other authors (Tribe, Zuckerman, Wasserman, Pundik) think that a conviction based on just statistical evidence is morally unacceptable. A few authors (Kaye, Stein) look at trial proceedings and their dialogical nature of accusation and defense. Finally, other authors (Tribe, Nesson) look at the larger socio-political picture. I think a lot can be learned from these different arguments. But besides the critics of statistical evidence and probability in the courtroom, there are the supporters: Kaplan, Finkelstein, Fairley, Lampert, Kaye, and many others I did not mentioned. The voice of the supporters should not be ignored. In order to move forward, what we now need is an approach that can integrate different dimensions—epistemic, procedural, and politico-moral—and an approach that can reconcile the supporter and critics of the use of statistics and probability in criminal trial. It is this conciliatory spirit that guided me in the earlier two chapters and that will guide me in the chapters that follow.

CHAPTER 4

HOW STATISTICS GET USED IN CRIMINAL CASES

What do I mean by 'statistical evidence'? As noted in chapter 1, it is difficult to offer a sharp definition, but we can identify a common pattern. The starting point consists in some statistical data, e.g., how many times a DNA profile shows up in a database; records about sales, travel, or hospital shifts. Next, from the data a statistical or probabilistic estimate is derived on the basis of a statistical model, e.g., an estimate of a DNA profile's frequency in a population is derived on the basis of a population model, or the probability of an event is derived from certain assumptions. Finally, the probabilistic and statistical estimates are used to draw inferences bearing on guilt or innocence.

In this chapter, I focus on three types of legal argument that rely heavily on statistical evidence. I call them the identification argument, the non-coincidence argument, and the total quantity argument. They rely on statistical evidence for three different purposes: identifying the perpetrator of a crime; assessing whether or not an event occurred accidentally; and finally, estimating a total quantify (e.g., the total amount of drug being illegally transported) when no direct measure of the total amount is available.

This chapter is best read together with chapter 5. Here I examine how statistics get used in criminal trials and there I examine how inferences can be drawn from them by using probability theory and Bayes' theorem.

4.1 STATISTICAL IDENTIFICATION

To begin with, consider what we might call a *statistical identification argument*, as follows:

> IDENTIFYING FEATURE. There is evidence that the perpetrator, or whoever visited the crime scene, possesses feature F.
>
> MATCH. The defendant possesses feature F.
>
> STATISTICAL FREQUENCY. Within a certain population of suspects, feature F occurs with a low frequency, e.g. 1 in 1 billion.
>
> PROBABILISTIC IDENTIFICATION. Hence, the perpetrator, or whoever visited the crime scene, must be the same person as the defendant, or more carefully, it is highly probable that they are the same person.

I will consider two criminal cases in which the prosecutor exploited the above reasoning pattern. Note that when it comes to eyewitness identification, the reasoning pattern is almost the same as the one above: the only difference is that the 'statistical frequency step' is absent. Although there is a rigourous way to formalize the statistical identification argument by means of Bayes' theorem, in this chapter I am interested in understanding how courts reason with statistical evidence.

4.1.1 Collins

On June 18, 1964, Juanita Brooks returned home from the grocery store. While she was standing outside her building, someone pushed her down to the ground. Shortly thereafter, she discovered that the money in her purse was missing. At the time of the incident, John Bass, who lived on a nearby street, heard a lot of crying and screaming, and he saw a woman run down the alley and enter a yellow convertible, which left promptly. Bass testified that a black male, with mustache and a beard, was at the wheel, and he described the woman as caucasian, with a dark blonde ponytail, wearing dark clothing.

The Los Angeles police, while investigating the robbery, came across Janet and Malcom Collins, who matched the description. They were placed under arrested, interrogated, and eventually tried for the robbery of Juanita Brooks. At trial, however, the prosecutor had difficulties in establishing the identity of the perpetrators: the victim had hardly seen Janet or Malcom, and the eyewitness testimony was found wanting because of inconsistencies with other evidence.[1] To bolster his case, the prosecutor introduced statistical evidence. A mathematics college instructor testified that, according to his calculations, the frequency of couples in California who would match the description was 1 in 12 million. As the frequency was so low, the prosecutor argued, the probability that the Collins were guilty was extremely high. The Collins were convicted, but on appeal, in the now celebrated case *People v. Collins* (1968), the California Supreme Court reversed the judgment.

While partly based on eyewitness testimony, the prosecutor's case relied heavily on statistical evidence. The court recognized three errors in the prosecutor's use of statistical evidence. First, the statistics presented at trial were mere guesses and lacked any scientific basis. This was a straightforward observation, based on elementary considerations from probability theory and statistics.[2] The second error was less straightforward. Even granting the correctness of the statistics, in order to find the Collins guilty beyond a reasonable doubt, the court noted, the jury should have assessed the accuracy of John Bass' testimony. After all, the victim's assailants might have not possessed the distinctive features the witness claimed they had. According to the court, the prosecutor's

[1] Bass testified that the day of the robbery Janet was wearing dark-colored clothing, but independent evidence showed that she was wearing light-colored clothing; also, some other evidence suggested that Malcom had shaved before June 18, the day of the robbery, but Bass testified that he had a beard.

[2] To calculate the frequency of the description, the mathematics college instructor did the following. First, he estimated the frequencies of single features, such as being a black male, having mustache, having a beard, etc. Next, the instructor applied the product rule and multiplied each frequency, thereby obtaining the frequency of the description. On the product rule, see chapter 5. The problem here is that the frequency of each features did not rest on any reliable body of data; and also, the product rule can be applied only if the frequencies of the features being multiplied are independent of each other.

emphasis on statistical evidence unduly distracted the jury from the task of evaluating the accuracy of the witness testimony. Finally, the third error was this. Even assuming that the statistics were correct and that the testimony was accurate, the court reasoned that the statistics would only establish how infrequently a couple with certain distinctive features could be found in California, whereas proving guilt beyond a reasonable doubt required one to establish that *only one* such couple existed in California. The court calculated that, if the frequency of a couple with the said characteristics was 1 in 12 million, there would be a 40 percent chance that a couple with such characteristics was unique in California, and such low chance was not enough to establish guilt beyond a reasonable doubt.[3]

On account of these three mistakes, the court ruled that the statistical evidence was inadmissible and *a fortiori* insufficient for a conviction. Interpreting *Collins* as a wholesale dismissal of the usefulness of statistical evidence for trial proceedings would be an exaggeration. The decision is better interpreted as providing a three-pronged test for whether statistical evidence (of a certain kind) is sufficient to justify a criminal conviction.[4] The court's test establishes that low-frequency statistics of distinctive characteristics that are shared by the defendant and the perpetrator, such as facial features and distinctive traits, are sufficient for a criminal conviction, provided:

(i) the statistics concerning the traits' frequencies are sound;[5]

(ii) no other substantial source of uncertainty is unaccounted for;[6]

(iii) the low frequencies establish the uniqueness of the traits.

Evidently, the statistics in *Collins* did not meet any of the three conditions. In fact, it is hard to see how any kind of statistical evidence could, because while the first two

[3] This number was calculated using the binomial distribution; see chapter 5.

[4] A piece of evidence is sufficient for a criminal conviction whenever a reasonable juror could find the defendant guilty beyond a reasonable doubt on the basis of the said piece of evidence. The standard of sufficiency is not the same as the actual satisfaction of the beyond a reasonable doubt standard: the fact that a reasonable juror could find someone guilty beyond a reasonable doubt does not entail that the juror necessarily would.

[5] The statistics should be arrived at using a sound methodology and data collection.

[6] In *Collins*, the other source of uncertainty that was not accounted for was the fallibility of the testimony by Bass.

conditions could be satisfied, the third is impossible to meet. No matter how low the frequencies of certain traits are, no judgment of uniqueness will be attainable on purely statistical grounds: at most, statistics can establish a high probability of uniqueness. As a result, a weaker third condition, and perhaps a more realistic one, would be as follows:

(iii*) frequencies are *low enough* to support the traits' quasi-uniqueness.

By "quasi-uniqueness" I mean that the traits in question are unique with a sufficiently high probability, but of course, the question now arises of what counts as an acceptable probability of uniqueness. In order to satisfy the beyond a reasonable doubt standard, how high should the resulting probability of traits' uniqueness be? The logic of the *Collins* decision, I think, leads us to ask this sort of questions. However natural they may be, questions about threshold probabilities have no answer, and in the dissertation I shall argue that the legal fact-finding process is at odds with any general pronouncement that a certain threshold probability (of traits' uniqueness or of guilt) is sufficient for a conviction.

4.1.2 Rush

It is instructive to turn to one of the most powerful forms of statistical evidence currently available, DNA evidence.[7] DNA evidence is a form of trace evidence, for it consists of traces found at the crime scene, e.g. bloodstains, hair, skin tissue, semen, etc. The traces are then analyzed in a laboratory and a DNA profile is created from them. A DNA profile is a codified representation of select portions of the human genome. The profile created from the traces is then compared to the DNA profile of the defendant. The goal is to find a genetic match. If a match is found, this does not yet mean that the defendant left the traces. This will depend on the frequency of the DNA profile for which the match was declared. The less frequent the profile, the more probative the match. In short, DNA evidence consists of two components: a laboratory test showing a genetic match; an estimate of the profile's statistical frequency. It is this second component that makes DNA evidence a form of statistical evidence.

[7] See (Kaye and Sensabaugh, 2000), (Kaye, 2010b) and (Wasserman, 2008).

We are now ready to examine a DNA case, which, I believe only apparently, offers a refutation of the proposition that a mere threshold probability, in itself, cannot be enough for a conviction. It is a growing phenomenon that individuals are tried and convicted on mostly statistical DNA evidence;[8] one such case is *People v. Rush* (1995).[9] The victim was robbed and raped. During a police lineup she identified Basheen Rush as the perpetrator, but at trial she was unable to identify him as her assailant. As a result, the case against Rush was mostly based on DNA evidence: laboratory testing showed that Rush's blood samples matched with the semen recovered from the victim, and that the frequency of the matching DNA profile was 1 in 500 million. Rush was convicted. On appeal, New York highest court upheld the decision, ruling that

> the testimony that there is a genetic match between the semen recovered from the victim of a rape and the blood of the defendant, a total stranger, and the statistical probability that anyone else was the source are 1 in 500 million is legally sufficient to support a guilty verdict.[10]

There is a striking similarity between the statistical evidence in *Collins* and in *Rush*. In both cases, the prosecutor established that the perpetrator and the defendant shared certain low-frequency characteristics;[11] and also, in both cases there was some uncertainty as to whether the perpetrator actually possessed the matching characteristics, although the uncertainty might have been more pronounced in *Collins*.[12]

[8] See (Roth, 2010).

[9] 630 N.Y.S.2d 631.

[10] *Rush* at 634. Notice that the court talked of a probability of 1 in 500 million, while I spoke of a frequency of 1 in 500 million. It is better to speak of an estimated frequency or a probability because no actual counting was done; rather, the estimate 1 in 500 million is the result of genetic population models. Besides, the court committed what is known as the prosecutor's fallacy because it talked about the probability that someone else was the source, whereas 1 in 500 million is the probability that someone who is not not source would be found with a matching DNA. The error here seems innocuous. For more details on this, see chapter 5.

[11] For the Collins, being a couple driving a yellow convertible, etc.; for Rush, possessing a certain DNA profile. The frequencies are, respectively, 1 in 12 million and 1 in 500 million

[12] In *Collins*, the uncertainty is due to the fallibility of the eyewitness testimony. In *Rush*, the uncertainty is due two factors: first, laboratory testing might have erred in declaring the match; second, the mere presence of DNA trace at the crime scene is not an indication of guilt, because the traces could have been left innocently.

The two cases are also different: the statistical frequencies are different—1 in 500 million *versus* 1 in 12 million—and allegedly they are much more dependable in the DNA case.[13] The other significant difference is that in *Collins* the prosecutor relied on eyewitness testimony to link the statistics to the crime scene, whereas in *Rush* forensic experts and laboratory analyses linked the DNA statistics to the crime by showing that the matching DNA traces were at the crime scene and could not have been left there innocently.

Rush is part of a family of court decisions suggesting that a conviction almost entirely based on DNA evidence statistics is warranted.[14] This is not in tension with *Collins* and its three-pronged test, at least if condition (iii) is interpreted in its weaker form. In fact, the statistical DNA evidence in *Rush* met condition (i); it arguably satisfied condition (ii), as well, although the court paid relatively little attention to this issue; and finally, proviso (iii*) was met because, over a population of 60 million people, the probability that a matching DNA with a frequency of 1 in 500 million is unique equals 0.95.[15] This means that the court in *Rush* may be interpreted as *implicitly admitting* that when the probability of uniqueness is sufficiently high, e.g. 0.95, DNA statistics alone warrant a conviction.

But we should be careful not to take the New York Court of Appeals to say what it did not say. The Court said that a DNA profile frequency of 1 in 500 million, given particular circumstances (i.e. rape case; semen taken from the victim and matching the defendant's blood; defendant being a total stranger to the victim), was enough to establish guilt beyond a reasonable doubt. Interestingly enough, the *Rush* decision did not comment on the uniqueness of the matching characteristics (the DNA profile, in this case), which instead played an important role in *Collins*. The *Rush* court simply stated that a low enough frequency of the DNA profile can satisfy the criminal standard; it did not require that the uniqueness, or quasi-uniqueness, of the DNA profile be proven.[16]

[13] DNA statistics, especially when they are so extreme, may not be dependable; see (Buckleton, 2005b).

[14] See (Roth, 2010).

[15] The value was calculated using the formula for the *binomial distribution*; see chapter 5.

[16] Instead of focusing the uniqueness of the DNA profile, it might be best to estimate the probability that the defendant is the source. For a comparison of the uniqueness approach with other methods, see

It is not premature to say that courts now deem statistical DNA evidence potentially sufficient for a criminal conviction. Once the errors identified in *Collins* are avoided, many appellate courts today are willing to uphold convictions based on mostly DNA statistical evidence. In the dissertation I argue that this position is justified, yet we should not expect from the courts, nor from legal scholars, a statement of a threshold probability of guilt as the minimum quantitative value that is sufficient for a criminal conviction. As remarked before, this would be at odds with the logic of the legal fact-finding process.

4.2 Coincidence or purposeful conduct?

Criminal cases often contain what we might call non-coincidence arguments. Here is a summary of a famous British case, Rex v. Smith, 11 Cr. App. R. 229 (1915), involving one such argument:

> George Joseph Smith was accused of drowning Bessie Mundy in the small bathtub of their quarters in a boarding house. Mundy had left all her property to Smith in a will executed after their 'marriage' (Smith was already married). The trial court allowed the prosecution to prove the deaths of two other women who had gone through marriage ceremonies with Smith and to argue that the circumstances surrounding their deaths in their bathtubs were remarkably similar ... The Court of Criminal Appeal affirmed the resulting conviction on the ground that the evidence in connection with Mundy's death alone made out a prima facie case. (Feinberg and Kaye, 1991, p. 63)

Intuitively, it seems that if not only Ms. Bessie, but two other women drowned in the presence of Mr. Smith, the probability that the drawing of Ms. Bessie happened by chance decreases. The argument can be schematized, as follows:

> OCCURRENCE. Event E occurred in the presence of defendant D.
>
> SIMILARITY. Other events E^{*}, which are all similar to E, occurred in the presence of defendant D.

chapter 5.

> STATISTICAL FREQUENCY. On the hypothesis that everything happened by chance, the series of similar events consisting of E and E^*, all occurring in the presence of D, is statistically very infrequent.
>
> NON-COINCIDENCE. Event E could not have occurred as a result of chance; rather it was—most probably—the result of D's purposeful conduct.

The existence of similar events, which all occurred in the presence of defendant D, plays a key role in the arguent, at least rhetorically. Consider now the same argument without the similarity step:

> OCCURRENCE. Event E occurred.
>
> STATISTICAL FREQUENCY. On the hypothesis that everything happened by chance, events such as E, occurring in the presence of D, are statistically very infrequent.
>
> NON-COINCIDENCE. Event E could not have occurred as a result of chance; rather it was—most probably—the result of D's purposeful conduct.

The argument now seems plainly fallacious. At best, it is like a statistical or probabilistic modus tollens: if the hypothesis of chance is true, then E is unlikely to be true; but E is true; therefore, the hypothesis of chance is unlikely to be true. Suppose you and a friend of yours buy one lottery ticket each. As it turns out, you lose the lottery but—with much surprise—your friend wins. Winning the lottery is something that occurs very rarely. Not everybody wins the lottery, only very few people do. Yet, although your friend won, it makes no sense to conclude that she rigged the lottery. After all, incredible coincidences are bound to happen. Someone is bound to win the lottery, although it would be an incredible coincidence if the winner was you or your friend. So, the above argument is fallacious. But if it is, does the addition of the similarity step make much of a difference? If your friend wins not one, but three consecutive lotteries, you might be suspicious that she rigged the lottery. Is that really so? There are people who won more than one lottery, after all. (A formal method to address this issue is given in chapter 5.)

All in all, statistical non-coincidence arguments, when they are not downright fallacious, have to be formulated with extreme care. We will see two examples in which such arguments where formulated without much care, and what followed—unfortunately—were gross miscarriages of justice.[17]

4.2.1 Sally Clark

I begin with the much discussed British case *R v. Clark* (1999). Sally Clark's babies, a few months old, both died in the presence of their mother. The first baby died in 1996 and the second in 1998. Sally was arrested and charged with murdering both her babies. At trial, the pediatrician Roy Meadow testified that the probability of two consecutive cot deaths, in a family similar to the Clarks, was 1 in 73 million. Theses statistics were thought to be impressive enough to rule out the possibility that the two deaths occurred accidentally or because of natural causes. Sally was convicted to life imprisonment in 1999 and the conviction was upheld on appeal in 2000.[18] On the second appeal, in 2003, Sally was released in light of new medical evidence but also on the grounds that the statistics were flawed and unreliable.[19] As a result, professor Meadow incurred several disciplinary sanctions for profesional misconduct and some 250 similar cases involving cot deaths were reopened.[20]

Just as with the first problem identified by the *Collins* court, Meadow's statistical calculations were flawed. He had no basis to give the numbers he gave. The undeniable yet often overlooked moral is that statistical data should be supplied to the courts with extreme care and by highly qualified professionals. But besides this, what if the statistics were sound? Would a conviction against Sally Clark be warranted then? Some have noted that, even if the figure of 1 in 73 million was correct, this would simply show that the probability of accidental death was very small. Yet, this probability is of little

[17] Here is a case worth mentioning, at least in a footnote. John Veysey had his house on fire four times in a relatively short period of time and each time his insurance company paid. He was tried for fraud and arson. The incriminating evidence contained statistical evidence, showing that the estimated frequency of four residential fires occurring by chance during a certain period of time was only one in 1.773 trillion. This evidence was used by the prosecutor to claim that Veysey's houses being on fire was not an accident, but the result of intentional conduct. See *United States v. Veysey*, 334 F.3d 600 (7th Circuit, 2003).

[18] See R v. Clark, EWCA Crim 54 (2nd October, 2000).

[19] See R v. Clark, EWCA Crim 1020 (11th April, 2003).

[20] See the BBC on-line article: *news.bbc.co.uk/2/hi/uk_news/3412307.stm*.

significance unless it is compared to another probability, that of a mother killing both her babies intentionally. This second probability, if it were in the vicinity of 1 in 73 million, would drastically undermine the significance of the first probability.[21] After all, both an accidental death and a mother's intentional killing of her babies are extremely unlikely hypotheses. The importance of contrasting competing hypotheses suggests that not only is it important to begin with sound statistical data, but also that these data should be interpreted and assessed properly.

4.2.2 Lucia de Berk

Lucia de Berk is a licensed pediatric nurse from the Hague in the Netherlands. On the morning of September 4, 2001, at Juliana's children Hospital in the Hague, a baby died unexpectedly. The nurse in charge when the baby died was Lucia de Berk. On the day after the incident, another nurse went to talk to her superior and reported that Lucia had been present at too many resuscitations. The director of the hospital later found out that, indeed, Lucia had been involved in an unreasonably high number of incidents. The director shared his findings with the police, and on December 13, 2001, Lucia was arrested and charged with multiple murders.

Let's take a look at the numbers:

Shifts at Juliana's hosp.	without incidents	with incident
without Lucia	887	0
with Lucia	134	8

The numbers look striking: Lucia was present during *all* incidents! It is very tempting to conclude that this cannot be a coincidence, and that Lucia must have done some-

[21] For review of the literature, see (Galavotti, 2012).

thing. This was essentially the argument made by the expert witness for the prosecutor, Dr. Elffers, a law professor with an undergraduate degree in statistics. He calculated the conditional probability that all incidents occurred during Lucia's shift, given the total number of incidents and the total number of shifts. He used the formula for the *hypergeometric distribution*, where n is the total number of shifts, r is the number of Lucia's shifts, x is the number of incidents during Lucia's shifts, k is the total number of incidents, and p is the probability of one incident occurring, as follows:

$$\frac{\binom{r}{x}p^x(1-p)^{r-x}\binom{n-r}{k-x}p^{k-x}(1-p)^{n-r-k+x}}{\binom{n}{k}p^k(1-p)^{n-k}}$$

The numerator is the probability that Lucia witnessed an x number of incidents, during a total number of r shifts, and that the other nurses witnessed a $k-x$ number of incidents over a total of $n-r$ shifts. The denominator is the probability that a k number of incidents happened. So, the above formula describes the conditional probability that Lucia witnessed the number of deaths she witnessed, given how many deaths occurred overall. With the data in the above table and with other data regarding Lucia's shifts in two other hospitals, Elffers' estimate of such probability was 1 in 342 million. What does this number mean? Does it mean that it is unlikely that the deaths during Lucia's shifts occurred as a result of mere chance? This seemed the interpretation of the trial court, which wrote:

> The court is of the opinion that the probabilistic calculations given by Dr H. Elffers in his report of May 29, 2002, entail that it must be considered extremely improbable that the suspect experienced all incidents mentioned in the indictment coincidentally. These calculations consequently show that it is highly probable that there is a connection between the presence of the suspect and the occurrence of an incident. Quoted from (Meester et al., 2007, p. 241)

The figure '1 in 342 million' was enough to convince the Court that Lucia de Berk killed the baby during her shift intentionally. Lucia was finally convicted on March 24, 2003 and sentenced to life imprisonment. The sentence was upheld on appeal, and only overturned by the Netherlands Court of Cassation (Schneps and Colmez, 2013).

Elffers' calculations have been contested on many counts. For instance, he assumed that the probability that an incident occurred does not depend on atmospheric conditions, the time of day, etc. He assumed that all nurses must have exactly the same probability of witnessing an incident. He also assumed that the occurrence of one death is independent of the occurrence of another. As Meester et al. (2007) have demonstrated at length, Elffers' probability model was overly simplistic; it made too many simplifying assumptions.

Another problem was that the trial court committed what is known as the *prosecutor's fallacy*. (More on this in chapter 5.) Even assuming that Elffers' figure was dependable, the estimated probability of 1 in 342 million is not the probability that Lucia is innocent or that the babies died accidentally. Rather, such a figure represents the probability that, given the hypothesis that everything happened by chance (call it C), we would witness as many deaths as those we witnessed during Lucia's shifts (call it D). In other words, Elffers' figure gives us the probability of D given C, not the probability of C given D. Yet, the Court took Elffers' figure to give us the probability of C given D. This was a mistake that led to a miscarriage of justice to the detriment of Lucia de Berk.

***

In both the Sally Clark case and the Lucia de Berk case, two gross errors were committed. The probabilistic or statistical estimates drawn from the data were incorrect or based on statistical models that were too simplistic. Second, even if these estimates were trustworthy, they were misinterpreted and hasty inferences were drawn from them. But now, for the sake of argument, let us suppose that the statistical estimates were trustworthy and properly interpreted, so that the probability of Sally intentionally killing her babies would turn out to be extremely high, and the same with Lucia de Berk. On this, admittedly far-fetched variation of our two cases, we may now ask whether a high probability is enough to warrant the murder conviction of Sally Clark and of Lucia de Berk. Consistently with DNA cases, we may be tempted to answer that it is enough. On the other hand, the hypothetical scenario *Prisoners* from chapter 1 suggests the opposite conclusion. The important question here is whether our modified versions of the Sally Clark case and of the Lucia de Berk case are more similar to *Rush* or more similar to *Prisoners*. In all cases, the available statistics are properly interpreted and they establish that the defendant's guilt is highly probable. Is DNA evidence a better example of statistics-based evidence? Is the difference that the guilt probability on the basis of DNA evidence can be impressively high, or is there more to be said? I think the difference does not lie in an impressively high guilt probability. As I argue in the rest of the dissertation, the difference has to do with whether or not the prosecutor has presented a well-specified case against the defendant, and with whether or not the prosecutor's case can withstand challenges. More on this in chapters 7 and 8.

4.3 ESTIMATING QUANTITIES

Sometimes an individual is found—*in flagrante delicto*—stealing a certain amount of money or carrying a certain amount of illegal substance. Suppose he goes to trial and he is convicted. Suppose, also, that further evidence suggests that the said individual stole more money or carried more drugs on other, undocumented occasions. For the purpose of sentencing or compensation, there is the need to calculate how much he stole or carried *in total*. Statistics can be used in these cases to estimate the total amount of drugs or money. This is what I call the *statistical total quantity argument*. We will now see two such examples.

4.3.1 Drug smuggling

Charles Shonubi flew from Nigeria to the United States. He was found carrying 427.4 grams of heroin at JFK airport in New York. He was placed under arrest, went to trial, and was convicted. At sentencing, federal district judge Weinstein concluded that Shonubi carried a total of 3,419.2 grams. The travel record showed that Shonubi made seven other trips between New York and Nigeria, so Weinstein multiplied the 427.4 gr. amount by the total number of trips. In accordance with the *Federal Sentencing Guidelines*, Shonubi was sentenced to 151 months in prison, a longer term than if he were found carrying only 427.4 grams.[22] The Court of Appeal disagreed with judge Weinstein and remanded for a resentence. The main complaint was that there was no "specific evidence" that Shonubi carried as much as 3,419.2 grams.[23]

On remand, Weinstein relied on data from the U.S. Custom Service about 117 other Nigerian drug smugglers who were found carrying heroin between September 1, 1990 and December 10, 1991. The data are reported in the table below:

Net weight (grams)	Number of Occurrences
0-100	1
100-200	7
200-300	13
300-400	32
400-500	31
500-600	21
600-700	6
700-800	1
800-900	2
900-1000	2
1000-1100	0
1100-1200	0
1200-1300	1

A cursory look at the data shows that in the overwhelming majority of cases, drug smugglers carry more than 100 grams per trip. So, a conservative estimate suggests that during his other seven trips, Shonubi carried a total of at least 700 grams. But what if Shonubi carried less on some of the trips?

Dr. Boyum, the expert witness for the prosecutor, looked at the data more aggressively, as it were. He used a Monte Carlo simulation and concluded that there was a 99 percent probability that Shonubi carried at least 2090.2 grams of heroin during his seven other trips. This is a disputable conclusion and it did not go unchallenged. The defense

[22] United States v. Shonubi, 802 F. Supp. 859 (E.D.N.Y., 1992).

[23] United States v. Shonubi, 998 F.2d 84 (2d Cir, 1993).

hired Dr. Finkelstein, a well-known scholar in the field of law and statistics (Finkelstein and Levin, 2001; Kaye, 1980; Finkelstein and Fairley, 1970). Finkelstein's principal objection was that Boyum failed to take into account the "intertrip variation." Due to a learning curve effect, it is plausible that most smugglers tend to carry smaller amounts during their earlier trips and bigger amounts during their later trips. Finkelstein thought that a regression analysis should have been used to identify a correlation between trip number and amount of drug carried. He also had reservations about using data about other drug smugglers for drawing inferences about Shonubi.

Judge Weinstein appointed a panel of experts, composed of Dr. Peter Tillers and David Shum, well-known scholars in the field of evidence law and legal reasoning. They both criticized Boyum's approach and suggested their own method, which was quite simple. In order to account for Finkelstein's intertrip effect, they looked for the smallest amount in the 117 data points available, i.e. 42.156 grams. They took this to be the smallest amount Shonubi carried, presumably on his first trip. Next, they took the 427.4 gr. amount on his eighth trip to be the largest Shonubi carried. What needed to be estimated, then, were the amounts carried between the first and eighth trip, beginning with 42.56 grams and reaching 427.4 grams. To this end, Tillers and Shum considered two possible rates of increase. On one prospected rate, the amounts grew as follows: 42.56; 60; 100; 125; 175; 225; 325; 527.4; 427.4. On the other, the amounts grew as follows; 42.56; 105; 160; 220; 225; 275; 325; 427.4. In either case, the total amount Shonubi carried exceeded 1,000 grams.

Tillers and Shum spent a significant amount of time learning about the socio-economics of drug smuggling. They learnt how smugglers carry the substance by wrapping it in small balloons and ingesting a variable number of them per trip. So, the two prospected rates of increase were based on domain-specific knowledge about drug smuggling. Judge Weinstein concluded that Shonubi carried at least 1,000 grams and he sentenced him accordingly.[24] The Court of Appeal disagreed again, lamenting that there was no specific evidence that Shonubi carried as much as 1,000 grams.[25]

[24] United States v. Shonubi, 895 F.Supp. 460 (E.D.N.Y., 1997).

[25] United States v. Shonubi, 103 F.3d 1085 (2d Cir., 1997).

The Shonubi case sparked controversy among legal scholars. Some, e.g. Tillers (1997), thought that the request for "specific evidence" was non-sensical and others, e.g. Colyvan et al. (2001), thought it more reasonable. Note that here the matter is complicated because we are dealing with a sentencing hearing in which the governing standard is not beyond a reasonable doubt, but a lower standard, such as clear and convincing evidence. This peculiarity is also what makes the case interesting. Even relative to a lower standard, some Courts are unwilling to reach a conclusion on the basis of a statistical model.

4.3.2 Theft

Varn Lombard was an employee in a grocery store. The owner, Don Erbert, suspected her of stealing during her shifts, so a video camera was installed in the store. On September 9, 2007, Lombard was found stealing $282 from the cash register. She went to trial, and a jury convicted her of burglary and grand theft. At the restitution hearing, the question emerged of how much Lombard stole in total, prior to the day she was caught. Erbert claimed he believed Lombard stole over $100,000 dollars. There was some circumstantial evidence suggesting that sales went up as soon as Lombard was fired and that sales were unreasonably low during Lombard's shifts. Further, a number of thefts occurred since Lombard started working for Erbert's store. More importantly, Erbert's estimate was based on a simple formula:

$$\textit{excessive no sale events} \times \textit{average amount per no sale event} \times \textit{number shifts}$$

Erbert, who had a background in statistics, randomly sampled 6 months of cash register tape. He identified Lombard's shifts and compared them to shifts by other employees during the same hours. He averaged the number of no sale entries during Lombard's shifts and compared them to the average number of no sales entries by other employees. He calculated an average of 10.3 excessive no sale events for Lombard. To calculate the average amount stolen per no sale event, Erbert divided $282 by 10.3, or to get a more conservative estimate, he divided $160 by 10.3. Finally, Erbert considered 674 shifts or more conservatively 523 shifts by Lombard. The time period was between October 2004 and August 2007, which is roughly the time between the first documented theft and when Lombard was found stealing. On one estimate, Lombard stole roughly $200,000 dollars; on the more conservative estimate, she stole roughly $70,000. The trial court was persuaded by Erbert's statistical argument. The Idaho Court of Appeals agreed:

> Considering all the evidence in the record, including the evidence in the statistical model presented by Erbert, we conclude that there was substantial evidence from which the district court could find that Mary Ann's Grocery suffered a loss of $85,000. State v. Lombard (2010), 149 Idaho 819. Petition for review was denied by the Supreme Court of Idaho.

We should be wary that at a restitution hearing the standard of proof is preponderance of the evidence, not beyond a reasonable doubt. It is not clear from the record, however, whether or not the $85,000 estimate of stolen money influenced the jury's finding of grand theft. Without the statistical estimate, the money stolen totaled only $286 dollars, a quantity insufficient to qualify as grand theft. Establishing grant theft requires a number of aggravating circumstances. The aggravating circumstance that seems to best apply to Lombard is the following:

> 7. When any series of thefts, comprised of individual thefts having a value of one thousand dollars ($1,000) or less, are part of a common scheme or plan, the thefts may be aggregated in one count and the sum of the value of all of the thefts shall be the value considered in determining whether the value exceeds one thousand dollars ($1,000). IDAHO STATUES, TITLE 18, CRIMES AND PUNISHMENTS, CHAPTER 24, THEFT.

So, the $85,000 estimated amount might have played a role in convincing the jury that Lombard was guilty of grand theft. Or at least, the estimated amount influenced the district court in determining the length of Lombard's sentence. Lombard's appellate brief, in fact, states that "[t]he consideration of these [statistical] models as a restitution amount influenced the District Court's decision on how long Verna Lombard's sentence would be with the District Court's intent that Verna be on probation for up to eight years in order to pay the restitution."

***

Both the Shonubi case and the Lombard case are procedurally similar. The inferences drawn from the statistical model are not used for the purpose of a criminal conviction; they are not used directly at trial to establish guilt beyond a reasonable doubt. Rather, they are used after the conviction, either at sentencing or at a restitution hearing. But notice that sentencing and restitution hearings can yield decisions that might be as damning for a defendant as a criminal conviction itself. In the Shonubi case, for instance, it makes a significant difference how much drug Shonubi carried in total. If the conclusion is that he carried more than 1,000 grams, the Federal Sentencing Guidelines mandate that he should be sentenced to at least 121 months in prison. If the conclusion is that he carried only up to 700 grams, he can be sentenced to at most 97 months. There is a a non-negligible difference of more than two years in prison. In the case of Lombard, the Court ruled that she should restitute as much as $ 85,000, but it would have ruled differently had it reached a different conclusion as to the total amount that Lombard stole.

4.4 What's at issue

This cursory review of the case law has demonstrated that the courts' attitudes toward statistical evidence vary on a case-by-case basis. Courts can express a strong resistance and be very critical (*Shonubi*, *Collins*); or they can look upon statistical evidence favorably (*Lombard*, *Rush*), even when it is flawed (*Clark*, *de Berk*).

What should courts and fact-finders do with statistics? Are statistics enough for a criminal conviction or not? As already anticipated in chapter 1, my answer is of the form—*it depends*. In DNA cases such as *Rush*, the crucial role that statistics play in supporting a conviction seems justified. When we turn to cases such as *Clark* and *de Berk*, we must recognize that they often involve an improper use of statistics and hasty inferences. But it is still worth asking whether a conviction would be justified had the statistics not been plagued with flaws and misinterpretations. Finally, another question concerns cases such as *Lombard* and *Shonubi*. Why in one case the appellate court dismissed the statistics and the other it did not? Is there a tension between the two cases or not?

The answers to these questions belong to the intricacies and complexities of each case. But, at least, as an external observer who wants to make sense of criminal trials, I will offer a rule of thumb or a framework, which can help us approach the question of when statistics alone are enough to convict. I will rely extensively on two criteria: (1) whether the prosecutor's story of the crime is sufficiently specific; (2) whether the defendant has an effective opportunity to challenge the prosecutor's case. My view is that so long as statistical evidence does not restrict (1) and (2), it should not give rise to any peculiar, statistics-related worry. I develop this framework in the rest of the dissertation, in particular in chapters 7 and 8.

CHAPTER 5

BAYES IN THE COURTROOM

As the previous chapter demonstrated, statistical evidence is introduced in criminal trials for at least three purposes. First, statistics can be used to answer identification questions, such as whether the defendant was the source of the crime traces or whether he had contact with the crime scene. DNA evidence and the statistical frequency of a DNA profile are a vivid example. Second, statistics can also be used to assess whether certain events or clusters of events (e.g. the same house on fire multiple times; multiple consecutive deaths) are the result of an accident or of purposeful conduct. Third, statistics can be used to indirectly estimate total quantities (e.g. quantities of drugs illegally transported or amounts of stolen money) when no direct measure of such total quantities is available. This chapter will illustrate how Bayes' rule can help us assess the probative value of statistical evidence in these three types of cases. The chapter also contains a basic introduction to probability theory, its mathematics and its philosophy, and to Bayes' rule.

5.1 THE MATHEMATICS OF PROBABILITY

I begin with the mathematical treatment of probability.[1] Suppose P is a function from a set of propositions into the real numbers; then, P is a probability function if it satisfies the Kolmogorov axioms below:

> NORMALITY: $0 \leq P(A) \leq 1$, for any proposition A;
>
> CERTAINTY: $P(\top) = 1$, with $\top$ any logical tautology; and
>
> ADDITIVITY: $P(A \vee B) = P(A) + P(B)$, with A, B mutually inconsistent propositions.

In other words, a probability function assigns to every proposition a real number between 0 and 1. It assigns the value 1 to tautologies. And to the disjunction of inconsistent propositions, it assigns the sum of their probabilities. Now that we have defined $P(A)$, we can define the *conditional probability* of A given some other proposition B, as follows:

> CONDITIONAL PROBABILITY: $P(A|B) = \frac{P(A \wedge B)}{P(B)}$.

Some authors take conditional probability to be their starting point and formulate the axioms in terms of conditional probabilities. This makes things more complicated mathematically, but it is more appropriate because any probability estimate, after all, is conditional (Hájek, 2007).

Finally, some theorems can now be proven, such as:

> OVERLAP: $P(A \vee B) = [P(A) + P(B)] - P(A \wedge B)$, if A and B are not mutually exclusive;
>
> LOGICAL CONSEQUENCE: If A implies B, then $P(A) \leq P(B)$;
>
> NEGATION: $P(\neg A) = 1 - P(A)$; and
>
> TOTAL PROBABILITY: $P(A) = P(A|B)P(B) + P(A|\neg B)P(\neg B)$.

[1] For an introduction to probability, see (Skyrms, 1999) and (Hacking, 2001).

It is worth distinguishing probability theory from its underlying logic. Every formulation of a theory of probability, such as the one above, must rest on an underlying logic, typically classical logic (or alternatively, set theory). Some theorems fail if the underlying logic is non-classical. For instance, to prove the negation rule, one begins with $P(A \vee \neg A) = 1$, which holds because $A \vee \neg A$ is a classical tautology. It then follows by additivity that $P(A) + P(\neg A) = 1$, whence $P(\neg A) = 1 - P(A)$. The proof gets off the ground because $A \vee \neg A$ is a classical tautology. If the underlying logic were intuitionistic, $A \vee \neg A$ would not be a tautology, and thus the negation rule would not be a theorem.[2]

The negation rule has been the target of some criticism. Suppose we have little or no information about A and little or no information about the negation of A. Intuitively, the probability of both A and its negation should be low, for we have little information about them (Shafer, 1976). But probability theory forces us to assign a non-zero probability to A and its negation, while in fact we have little or no information whatsoever about them. Probability theory, as it were, *creates evidence from ignorance* (Cohen, 1977, 1981b). Yet, the problem need not be with probability theory itself. It might have more to do with the underlying logic which is assumed to be classical. We can, in fact, maintain probability theory without the rule for negation if we assume that the underlying logic is intuitionistic (Weatherson, 2003).

This intuitionistic turn of probability theory seems particularly well-suited for trial proceedings. Having little evidence that the defendant is guilty translates into a low probability of guilt. Yet, it would be strange to say that, as a consequence, the probability of innocence must be high, or at least it would be strange if a high probability of innocence is taken to mean that the evidence strongly supports a finding of innocence. More needs to be said, but I shall say no more about this contentious issue. These brief

[2] Similarly, to prove the rule of total probability, one relies on $P(A) = P(A \wedge B) + P(A \wedge \neg B)$, and since $P(A \wedge B) = P(A|B)P(B)$ and $P(A \wedge \neg B) = P(A|\neg B)P(\neg B)$, the theorem follows immediately. But the equivalence $P(A) = P(A \wedge B) + P(A \wedge \neg B)$ holds because, first, A and $(A \wedge B) \vee (A \wedge \neg B)$ are (classically) logically equivalent, so they must have the same probability; and second, $A \wedge B$ and $A \wedge \neg B$ are inconsistent so the probability of their sum must the same as the probability of their disjunction. Yet, notice that in intuitionistic logic A can be true without neither $A \wedge B$ nor $A \wedge \neg B$ being true.

remarks were meant to point out that we need not abandon probability theory to accomodate certain intuitions we have, as some have suggested that we do; changing the underlying logic might be enough. It is not for me to say here how difficult or mathematically costly this change might turn out to be.

5.2 INTERPRETATIONS OF PROBABILITY

The mathematical treatment of probability is neutral as to what probability values express or what they mean. I shall give here a brief overview of the main interpretations.[3] The most natural interpretation of probability is the *classical* interpretation. A statement of it is given by Laplace (1814) as follows:

> The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability.

What does it means to say that some cases are equally possible or equally probable? Laplace takes the notion of equi-possibility or equiprobability as primitive, or at best he says that two cases are equally probable when we are 'equally undecided about their existence.' In developing this idea, we can say that two cases are equally probable if the evidence in favour of (or against) each of them is perfectly balanced. (Incidentally, note the ambiguity here between 'we lack evidence for and against proposition A' and 'we have equally strong evidence for and against A.' In both cases, the evidence is perfectly balanced although for very different reasons: lack of evidence in one case and equally strong evidence in the other.) We can also say that two cases are equally probable if they are physically symmetric. Be that as it may, spelling out the notion of equiprobability requires us to invoke a *principle of indifference*, in an epistemological or physical sense.

[3] For introduction to the philosophy of probability, see (Galavotti, 2005) and (Hájek, 2010).

Moving away from the classical interpretation, some theorists believe that probabilities are *objective* features of the world. In particular, some authors think that probabilities apply to classes of events. They equate probabilities to *relative frequencies* in the case of finite classes of events, or to *limiting relative frequencies* in the case of infinite classes of random events, known as collectives (Reichenbach, 1949; von Mises, 1957). The frequentist interpretation of probability makes good sense for events that can be repeated in the long run, but it is at odds with probabilities assigned to single-case events.

Other theorists hold that probabilities are not objective, but *subjective* or more generally *epistemic*. Loosely put, the idea is that the probability of a proposition corresponds to an *agent's degrees of belief* in a proposition.[4] In an attempt to spell this out, one can say that degrees of belief in a proposition mirror the strength of one's evidence for that proposition, or more precisely, they mirror what an ideally rational agent takes to be the strength of one's evidence. What does that mean exactly? Famously, de Finetti (1937) wrote:

> Let us suppose that an individual is obliged to evaluate the rate p at which he would be ready to exchange the possession of an arbitrary sum S (positive or negative) dependent on the occurrence of a given event E, for the possession of the sum pS; we will say by definition that this number p is the measure of the degree of probability attributed by the individual considered to the event E.

More succinctly, p is S's degree of belief in event E if S is willing to pay p units of utility for a bet that pays one unit of utility if E, and 0 if not-E. For example, if S is willing to pay 0.9 dollars for a bet on E which would give S one dollar if E occurs, then according to de Finetti, S would have a 0.9 degree of credence in E. This is an operational definition of probability that reduces probabilities to an agent's betting behaviour. Unsurprisingly, it inherits some of the problems of behaviourism, for the agent might have an interest in misrepresenting his actual betting preferences. So, instead of speaking of *actual betting behaviour*, it is best to speak of *betting rates that an ideally*

[4] On degrees of credence, see (Erikson and Hájek, 2007).

rational agent would regard as fair. What is a fair bet? A fair bet is one for which the agent is indifferent between taking either side. Interestingly enough, talk of fair bets parallels the appeal to the principle of indifference in the classical interpretation.

De Finetti derived probabilities from one's betting preferences and utility estimates. Ramsey (1931) managed to derive probabilities and utilities from one's preferences alone.[5] In his representation theorem, Ramsey showed that, given certain axioms that the preferences of a rational agent should conform with, the agent can be represented as maximizing expected utility, where the latter is calculated as the product of a utility function and a probability function. In other words, probabilities (i.e. a rational agent's degrees of belief) are for Ramsey whatever is multiplied with utilities to yield expected utility.

Some claim that de Finetti and Ramsey get the order of explanation backwards. They derive probabilities from utilities and preferences, while it seems that it should be the other way around:

> [P]reference-based accounts of degree of belief ... get the order of explanation wrong ... Why do you prefer buying fire insurance for your house rather than leaving it uninsured? Because you have some small-but-non-negligible credence that your house will burn down, some large disutility attached to it doing so, and some smaller disutility to paying the insurance premium, such that the expectation calculations favor your getting insurance. ...... The order of explanation, or of prediction, is not: These are your betting dispositions, or preferences; thus, those are your credences and desirabilities.' Rather, it is: 'These are your credences and desirabilities; thus, those are your betting dispositions, or preferences.' (Erikson and Hájek, 2007, p. 207-8)

The issue is controversial. The formula for the expected utility of an action—where $A(x)$ denotes the consequences that will result from doing action A, and where EU,

[5] The key idea is that of an *ethically neutral* proposition—a proposition whose truth or falsity does not matter for the agent. There are presumably plenty of such propositions: the coin landing heads (or tails); the lottery ticket number 5758696 being drawn (unless it's ours); etc. Let h be an ethically neutral proposition, and let A be an outcome that the agent prefers to outcome B. We say that h has a probability of $1/2$ if the agent is indifferent between [A if h; B if non-h] and [B if h; A if non-h].

P, and u are abbreviations for expected utility, probability, and utility functions—is as follows:

$$EU(A) = \sum_{x \in X} P(A(x)) \times u(A(x)).$$

On a naive reading of it, the formula suggests that utilities and probabilities are the primitive notions, while expected utilities are the derivative ones. On the other hand, both de Finetti and Ramsey wanted to eliminate such an intuitive understanding of probabilities, and they suggested that we cannot understand probabilities independently of an agent's preferences over expected utilities.

I wish to mention a fourth interpretation of probability, which may not count as an interpretation in its own right. The statistician David Freedman wondered what it means to say e.g. that the probability of rain tomorrow is 0.6 or that the probability of an earthquake in California in the next thirty years is 0.8. Here is his answer:

> Another interpretation of probability seems useful for making earthquake predictions: probability is just the property of a mathematical model intended to describe some features of the natural world. For the model to be useful, it must be shown to be in good correspondence with the system it describes. That is where the science comes in. (Freedman, 2003, p. 5)

This interpretation—that probability is a property of a (probabilistic) model of the world—can be applied to a variety of settings. We construct models of fair lotteries and we assign probabilities accordingly. We construct models of physical, meteorological, and geological phenomena, and we assign probabilities accordingly. Now, I do not think this interpretation is in tension with the epistemic or subjective interpretation. In fact, an ideally rational agent who has certain betting dispositions, or who considers certain betting rates as fair, presumably does so on the basis of the information he has available, and such information can also be encoded (but need not only be encoded) in terms of scientific models of a situation.

I should stop here. I cannot enter into a very difficult debate regarding the foundations and the meaning of probability. Returning now to the matter at hand, after a brief excursus into the interpretation of probability, we should ask: What interpretation is best suited for the legal domain? Legal proceedings investigate past events that have already happened, events that are non-repeatable and that can be hardly located in a set of equiprobable cases. Thus, both the classical and the frequentist interpretation seem ill-suited for the legal domain. What remains is the epistemic interpretation in terms of degrees of belief. Unsurprisingly, many legal probabilists have explicitly endorsed the epistemic interpretation (Kaplan, 1968; Finkelstein and Fairley, 1970; Cullison, 1969; Robertson and Vignaux, 1993).

The open question, then, is how we should interpret degrees of belief if we are interested in the probability of guilt or in the probability of propositions that are relevant to the issue of guilt. One way to do so, following de Finetti, is to interpret them as fair betting rates on a proposition, where such betting rates are entertained by an ideally rational agent. The assumption is that the hypothetical agent has access to all the evidence available at trial and has a perfect ability to assess its probative value. For those who are allergic to the betting interpretation, we can simply say, more flatfootedly, that that one's degrees of belief in a proposition must be commensurate with the evidence presented at trial (and absence thereof). The further constraint on one's degrees of belief is *coherence*, i.e. compliance with the axioms of probability, Bayes' rule and Bayes' updating if probability. To the last two, I now turn.

5.3 Bayes' rule and Bayes' updating

Suppose you formulated a hypothesis H and have some evidence E for it, and now you want to know the probability of H given E. Bayes' rule gives you the answer, as follows:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\neg H)P(\neg H)}.^{6}$$

Bayes' rule allows to calculate the probability of H given E from three other items: (i) the probability of H regardless of E; (ii) the probability of the evidence $P(E)$ which, by

[6] For an introduction to the many applications of Bayes' rule, see (Swinburne, 2002).

the rule of total evidence, equals $P(E|H)P(H)+P(E|\neg H)P(\neg H)$; (iii) the likelihood $P(E|H)$, i.e. the probability of E given H. Probability $P(H)$ measures the probability of H before taking into account the new piece of evidence E. Probability $P(E)$ is a measure of how unusual or surprising it is to obtain E. Finally, likelihood $P(E|H)$ is a measure of how much the hypothesis H "accounts for" or "predicts" E.

(The term 'likelihood' is an unfortunate word choice, but one that is now universally made in the literature. While in common parlance 'probability' and 'likelihood' are interchangeable terms, within the Bayesian framework the term 'likelihood' designates the probability of the evidence given the hypothesis, $P(E|H)$, as opposed to the probability of the hypothesis given the evidence $P(H|H)$.)

There is another formulation of Bayes' rule, one that makes calculations easier. It is in terms of odds and I will use it extensively in giving a Bayesian analysis of DNA evidence:

$$\frac{P(H|E)}{P(\neg H|E)} = \frac{P(E|H)}{P(E|\neg H)} \times \frac{P(H)}{P(\neg H)}.$$

In other words,

$$\textit{posterior odds} = \textit{likelihood ratio} \times \textit{prior odds}.$$

The posterior probability $P(H|E)$ is usually given by $\frac{PO}{1+PO}$, were PO are the posterior odds.[7]

A couple of complications bear mentioning here. First, Bayes' rule is different from *Bayes' update*. The former is a mathematical theorem, while the latter is an epistemological thesis regarding how agents should update their probability assignments for a proposition over time in light of new evidence. Bayes' update says:

$$\textit{If you learn } E \textit{ at time } t_1 > t_0, \textit{ then } P_1(H) = P_0(H|E).$$

The idea is that at time t_0 an agent assigns a probability $P_0(H)$ to H; this is called the

[7] The pages that follow have benefited from reading (Dawid, 2002), (Thompson et al., 2003), and (Buckleton, 2005a).

prior probability. As the agent learns E at time t_1, Bayes' update says that the agent should equate the *posterior probability* of H at time t_1 with the probability of H given E at time t_0, namely $P_1(H) = P_0(H|E)$.

Another complication is that, in the formulation of Bayes' update, it is as though the proposition E was indisputably true. But cannot we be mistaken about E? Maybe it looks as though the evidence we have is E, but as a matter of fact the evidence is more properly described by another proposition E'. Bayes' update ignores the possibility that one might be mistaken about E. In other words, the assumption is that once the evidence E has been collected, the subject must have a degree of credence in E that equals one, so that if the evidence E is collected at time t_1, $P_1(E) = 1$. Some authors disagree and have suggested an alternative form of update, e.g. *Jeffrey's update*.

5.4 TAXI CABS

Let us know see an example of how to use Bayes' rule. Imagine that there are two taxi companies, Green Cabs Inc. and Blue Cabs Inc., whose vehicles are respectively painted green and blue. Green Cabs Inc. covers 85 percent of the market and Blue Cabs Inc. covers the rest. There are no other taxi companies around. On a misty day a cab hits and injures a passerby, but it drives off. A witness reports that it was a blue cab. The witness is right only 80 percent of the time. This means that his *reliability* equals 0.8 in the sense that he gets the color right 80 percent of the time. Given the witness report, what is the probability that the taxi cab involved in the accident was in fact blue? Let G be the hypothesis that the taxi was green and B the hypothesis that the taxi was blue; let W_b be our evidence, namely the witness reporting that the taxi was blue. The problem is to determine the value of $P(B|W_b)$.

To apply Bayes' rule, we need three items: the probability of the hypothesis regardless of the evidence; the likelihood; and the probability of the evidence. First, Green Cabs Inc. covers 85 percent of the market, so the probability $P(G)$, without taking into consideration further evidence, equals 0.85. Notice that $P(G) = 1 - P(B)$, so $P(B)$ equals 0.15. Second, we know that the witness is correct 80 percent of the time. This measures his reliability: when he sees a taxi, he can identify the color correctly 80 percent of the time. Thus, the likelihood $P(W_b|B)$ equals 0.8; also, $P(W_b|G)$ equals 0.2 because the witness is wrong 20 percent of the time.[8] Finally, what is the probability of the evidence $P(W_b)$? By the rule of total evidence, we have that $P(W_b) = P(W_b|B)P(B) + P(W_b|G)P(G)$. That is, $P(W_b) = 0.8 * 0.15 + 0.2 * 0.85 = 0.29$. We are now ready to apply Bayes' rule:

$$P(B|W_b) = \frac{P(W_b|B)P(B)}{P(W_b|B)P(B) + P(W_b|G)P(G)} = \frac{0.8 * 0.15}{0.29} =\approx 0.41.$$

The interesting result here is that, even if the witness is right 80 percent of the time, the probability that the taxi was in fact blue given the testimony is still quite low. The reason is that the probability of B regardless of the evidence is low and equals 0.15. The table below shows that by varying the probability of B we can arrive at different probabilities of B given the witness testimony, holding fixed the witness reliability:

$P(B)$	$P(G)$	$P(W_b|B)$	$P(B|W_b)$
0.15	0.85	0.8	0.41
0.25	0.75	0.8	0.57
0.35	0.65	0.8	0.68
0.45	0.55	0.8	0.76
0.50	0.50	0.8	0.80
0.55	0.45	0.8	0.83
0.65	0.35	0.8	0.88
0.75	0.25	0.8	0.92
0.85	0.15	0.8	0.95

Given Bayes' rule, people should assign a probability of 0.41 to B once they learn about W_b, or at least this is how people should update their probability assignment ac-

[8] The assumption is that the witness' reliability is 80 percent, i.e. the witness gets the color of the taxi right 80 percent of the time. Let W_g abbreviate 'the witness reports that the taxi was green.' We have $P(W_b|B) = P(W_g|G) = 0.8$. Now, since the witness can report that the taxi was blue or that it was green, W_g and W_b are exhaustive and mutually exclusive, whence $P(W_b|G) = 1 - P(W_g|G) = 0.2$.

cording to Bayes' update. Do people actually follow Bayes' rule and Bayes' update? They don't. Psychological experiments have shown that ordinary people, when confronted with scenarios such as Taxi Cabs, typically answer that the probability that the taxi was blue equals 0.8. Why? There are different explanations. One is that there is some ambiguity in the expression "the witness is correct 80 percent of the time." It might mean: when the witness sees a cab, he gets the colour right 80 percent of the time; that is, $P(W_b|B) = 0.8$ and $P(W_b|G) = 0.2$. But the expression could also be taken to mean: when the witness claims that the taxi is blue, he is correct 80 percent of the time; that is, $P(B|W_b) = 0.8$. If people adopt the second interpretation, their answer would naturally be that there is an 80 percent chance that the taxi was blue, given the witness report.

A second, related explanation is that people tend to confuse two different probabilities: $P(A|B)$ and $P(B|A)$. This tendency is know as *inversion fallacy*. In our example the confusion would be between $P(W_b|B)$ and $P(B|W_b)$. People might take both probabilities to be 0.8, though the former equals 0.8 and the latter 0.41 (at least, according to Bayes' rule). As the table indicates, $P(W_b|B)$ and $P(B|W_b)$ are interchangeable only if the prior probability of B equals 0.5. And, presumably, the reason why people confuse the two probabilities is that they disregard prior probabilities and only take into account the witness report. This tendency is known as *base rate neglect* (Kahneman and Tversky, 1982).

Some scholars have challenged the conclusion that 0.41 is the correct probability assignment. They have argued that ordinary people might not be committing a fallacy, after all (Cohen, 1981b; Koehler, 1996b). The problem boils down to the question: Why should our probability assignments (or our degrees of belief) conform to the Kolmogorov axioms of probability, Bayes' rule, and Bayes' update? A standard answer is that if one does otherwise, he would be subject to a *Dutch Book* or a *Diachronic Dutch Book*. This means that he would be willing to take bets in which he loses money no matter what happens. I do not enter into this debate here. Bayes' rule remains a very powerful and elegant tool, and I now turn to its applications to criminal trials.

5.5 IDENTIFICATION EVIDENCE

In criminal cases we wish to know whether the defendant is guilty or not. So, given a body of evidence E, we are interested in knowing the value of $P(G|E)$, and for that purpose, Bayes' rule can be helpful. This holds in theory, although in practice calculating the probability of guilt might be excessively complicated. Some authors have tried to show that this can be done (Schum and Kadane, 1996). Here I prefer to illustrate the applications of Bayes' rule in a more modest yet still interesting way. I will show that Bayes' rule can help us estimate the probability of intermediate propositions which are crucial to determining the defendant's guilt. An example is offered by *identification propositions*, such as whether the defendant was present at the crime scene, or whether he left certain traces found at the crime scene. Identification propositions are different from those concerning what the defendant was doing at the crime scene, or whether he harmed the victim intentionally or accidentally.

Let us consider one of the most powerful forms of statistical evidence currently available, DNA evidence (Kaye and Sensabaugh, 2000; Kaye, 2010b; Wasserman, 2008). DNA evidence is a form of trace evidence, for it consists of traces found at the crime scene, e.g. bloodstains, hair, skin tissues, semen, etc. The traces are then analyzed in a laboratory and a DNA profile is created from them. A DNA profile is a codified representation of select portions of the human genome. The profile created from the traces is then compared to the DNA profile of the defendant. The goal is to find a genetic match. If a match is found, the question is whether the defendant was in fact the source of the traces, and Bayes' rule helps us answer this question.

5.5.1 A simplified Bayesian analysis

Let S be the proposition that the defendant is the source of the crime traces; let M be the proposition that the defendant and the traces match; let f represent the frequency of the DNA profile in question. We want to know the probability of S given M. Bayes' rule can be used as follows:

$$\frac{P(S|M)}{P(\neg S|M)} = \frac{P(M|S)}{P(M|\neg S)} \times \frac{P(S)}{P(\neg S)}.$$

The prior probabilities $P(S)$ and $P(\neg S)$ are difficult to estimate. One proposal is that, for a suspect population of n individuals, $P(S) = 1/n$. This proposal assumes that each individual is equally likely to be the source. Is this plausible? And how large is the suspect population? The largest suspect population is the earth population, but it might be more appropriate to take the population of a town or a country (Robertson and Vignaux, 1995). Another approach consists in taking the smallest possible value for $P(S)$, maybe even smaller than $1/n$, with n the earth population. The rationale is that there are incredibly many ways in which one could have left traces at the crime scene, and these ways are more in number that the total earth population (Friedman, 2000). Still another proposal suggests that we consider an interval of prior probabilities and that we assess their impact on the posterior probability (Finkelstein and Fairley, 1970). This might be the most sensible approach to take. All in all, assigning prior probabilities is difficult and it is one of most contested issues in the application of Bayes' rule to criminal trials. On the other hand, we should always begin somewhere; any reasoning begins with an assumption, and the need to assign prior probabilities only makes this explicit.

Next, we need to determine the value of $P(M|S)$, which is typically set to one. The reason is that if the defendant actually left the traces, the laboratory analyses would show a match. This is a simplification, for the laboratory analyses might fail to show a match even though the defendant did leave the traces (what is called a *false negative*). I leave this complication aside for the moment and put $P(M|S) = 1$. Finally, we should determine the value of $P(M|\neg S)$. If the defendant is not the source, what is the probability that a match would turn up? This is typically taken to be equivalent to the frequency f of the DNA profile. Again, this is a simplification because the defendant might not possess the DNA profile in question yet laboratory analyses show that he does (what is called a *false positive*). Complications aside, we have:

$$\frac{P(S|M)}{P(\neg S|M)} = \frac{1}{f} \times \frac{P(S)}{P(\neg S)}.^{9}$$

It is now useful to give some numerical examples. With different prior probability assignments and different likelihood ratios, the posterior odds are as follows:

$P(S) \approx$ Prior Odds	Likelihood Ratios	Posterior Odds	$P(S\|M)$
1 in 10 million	100 million	10	0.9
1 in 100 million	100 million	1	0.5
1 in 1 billion	100 million	0.1	0.09
1 in 10 million	1 billion	100	0.99
1 in 100 million	1 billion	10	0.9
1 in 1 billion	1 billion	1	0.5
1 in 10 million	10 billion	1000	0.999
1 in 100 million	10 billion	100	0.99
1 in 1 billion	10 billion	10	0.9

The analysis I've outlined here can be extended to other forms of trace evidence. For example, it can be extended to fingerprints: instead of a genetic match, we will have a match between fingerprint profiles. The difference with DNA evidence is that we do not have statistical estimates of the frequency of fingerprint profiles (Zabell, 2005). Many believe they are unique, but no one has ever proven it. Besides fingerprints, the same analysis can be used for blood types, glass evidence, and any other type of trace evidence. But again, we do need an estimate of the frequency of the features for which a match is declared.

5.5.2 A more sophisticated Bayesian analysis

Thus far I used a simplified Bayesian formalization of DNA evidence. I shall now sketch a more sophisticated analysis. The likelihood ratio, in its simplified version, was taken

[9] This simple analysis can be found, among others, in (Dawid, 2002) and (Balding, 2005).

to be $1/f$, with f the frequency of the DNA profile. Part of the simplification here is that $P(M|\neg S) = f$, which does not take into account the possibility of a false positive in the laboratory report (i.e. the laboratory reports a match but in fact the two samples do not match). To take this possibility into account, instead of a match M, we should properly speak of a reported match M_r. Now observe that we can obtain a reported match M_r in two situations: when there is a true, factual match M_t or when there is a true, factual non-match $\neg M_t$. So, we have:

$$P(M_r|\neg S) = P(M_r|M_t)P(M_t|\neg S) + P(M_r|\neg M_t)P(\neg M_t|\neg S).^{10}$$

Now, $P(M_t|\neg S)$ equals f, and thus $P(\neg M_t|\neg S)$ equals $1 - f$. Also, $P(M_r|\neg M_t)$ describes the laboratory's false positive rate, abbreviated by FP. Finally, $P(M_r|M_t)$ describes the laboratory's true positive rate, or the inverse of the laboratory's false negative rate, abbreviated FN. So, we have:

$$P(M|\neg S) = [(1 - FN) \times f] + [FP \times (1 - f)].^{11}$$

The formula above shows that in order to properly estimate the value of $P(M_r|\neg S)$, we need a statistical estimate of the DNA profile's frequency as well as of the laboratory's error rates (false positive and negative rates). Unfortunately, laboratory error rates are seldom available, though Bayes' rule indicates that we need them to make a proper assessment.

Another simplification is that $P(M_r|S)$ was assumed to equal one. This does not take into consideration the possibility of a false negative. As above, we have:

$$P(M_r|S) = P(M_r|M_t)P(M_t|S) + P(M_r|\neg M_t)P(\neg M_t|S).$$

In the equation, $P(M_t|S) = 1$ and $P(\neg M_t|S) = 0$. Thus, $P(M_r|S) = 1 - FN$. In sum,

[10] For an explanation, see (Thompson et al., 2003).

[11] See (Thompson et al., 2003) and (Buckleton, 2005a).

a more careful statement of the likelihood ratio is as follows:

$$\frac{P(M_r|S)}{P(M_r|\neg S)} = \frac{1 - FN}{[(1 - FN) \times f] + [FP \times (1 - f)]}.$$

To get a sense of how error rates can impact the likelihood ratio, suppose a DNA profile has a frequency as low as of 1 in 1 billion. Suppose laboratory false positive and false negative rates equal 0.01. If the likelihood ratio is simply equivalent to $1/f$, then it would be 1 billion. With our new formula, instead, we have:

$$\frac{P(M_r|S)}{P(M_r|\neg S)} = \frac{1 - 0.01}{[(1 - 0.01) \times 0.000000001] + [0.01 \times (1 - 0.000000001)]} \approx 100.$$ [12]

Relatively small error rates can therefore significantly diminish the probative value of DNA evidence by dramatically reducing the likelihood ratio; in our case we went from one billion to only one hundred.[13] The impact of laboratory error rate is an indication that no matter how low a profile's frequency turns out to be, it will be of relatively little significance if the frequency is unaccompanied by an estimate of the laboratory error rates. Bayes' rule allows us to take this aspect into consideration.

Finally, I wish to mention two further complications, concerning genetic matches and profiles' frequencies. I have been speaking of reported matches between genetic profiles. As a matter of fact, laboratory results do not yield any reported match; rather, they yield a congruence result between two DNA profiles, one from the defendant and one from the crime traces. The two profiles are never perfectly identical; they are more or less congruent (Kaye, 1993). Instead of a reported match, what we really have are two pieces of evidence: one is the laboratory test showing that the defendant has the genetic profile d and the other is a laboratory test showing that the traces have the genetic profile

[12] The calculation is as follows:
$= \frac{1-0.01}{[(1-0.01)\times 0.000000001]+[0.01\times(1-0.000000001)]}$
$= \frac{0.99}{0.99\times 0.000000001+[0.01\times(0.999999999)]}$
$= \frac{0.99}{0.00000000099+0.00999999999}$
$= \frac{0.99}{0.01000000098} \approx 98 \approx 100$

[13] For a amore systematic discussion of this point, see Thompson et al. (2003) who show how even a small margin of laboratory error can substantially diminish the probative value of DNA evidence.

t; call them G_d and G_t. Instead of M, we can use the conjunction $G_d \wedge G_t$. The likelihood ratio can therefore written more explicitly as follows:

$$\frac{P(G_d \wedge G_t|S)}{P(G_d \wedge G_t|\neg S)}.$$

This new formulation might not change the actual calculations, but it does make the reasoning clearer (Robertson and Vignaux, 1995).

The other complication I wish to mention concerns the frequency of DNA profiles. I have been speaking as though DNA profiles were distributed across a suspect population with a fixed stable frequency, as though each individual had the same probability of been assigned a genetic profile. This is not true. DNA profiles represent part of our genetic make-up, which we receive through evolution and natural selection, not through a randomized process. Naturally enough, people who are related are more likely to share the same DNA profile than people who are unrelated (Buckleton, 2005b; Weir, 2007). This is why forensic experts should always qualify their statements by saying that their frequency estimates apply to a population of unrelated individuals. This variability in a profile's frequencies can be accommodated in the Bayesian framework in two ways: by changing the prior probability and by changing the profile's frequency. We could also carry out different calculations: one for a population of unrelated individuals and another for a population of related individuals.

5.5.3 Other approaches

To appreciate the power and elegance of the Bayesian analysis of DNA evidence, it is useful to compare it with other approaches.

Frequencies and uniqueness. Ordinary people are notoriously bad at applying Bayes' theorem, as we've seen in the the Taxi Cabs example. Some authors have suggested that a more natural way to handle probabilities is by *natural frequencies* (Gigerenzer et al., 1999). The idea is simple. Suppose a DNA profile has a frequency f of 1 in 10 million. Suppose also that the defendant possesses the DNA profile in question and that the initial suspect population consists of 100 million people. What is the probability that the defendant is the source? The methods based on natural frequencies has a simple answer. First, one should calculate how many people having the DNA profile in question there are in the suspect population, by multiplying f times m, with m the size of the suspect population. Second, if there are $n > 1$ such individuals with the profile, the probability that the defendant is the source will be $1/n$. In our example, the expectation is that there will be 10 individuals sharing the profile in question and thus the probability that the defendant is the source is 1 in 10. The method works easily if $n > 1$, yet it is less clear what to do when $n < 1$.

When frequencies are particularly small, we are better off asking whether or not the DNA profile is unique, or whether there are other individual with the same DNA profile besides the defendant. I call this the *uniqueness method* (Mortera and Dawid, 2007). The formula of the *binomial distribution* allows us to calculate the probability that an event x—e.g. having a DNA profile—will show up a k number of times, over an s number of repetitions, where the event x has frequency f. For our purposes, suppose a DNA profile has a frequency f. Suppose also the size of the suspect population equals m and that the defendant possesses the profile in question. We want to know how probable it is that exactly one individual has the profile, given that at least one individual, the defendant, has it. Let n be a function measuring the number of people possessing the profile; we have:

$$P(n = 1|n \geq 1) = \frac{m \times f \times (1 - f)^{m-1}}{1 - (1 - f)^{m}}$$

The denominator describes the inverse of the probability that no one has the profile in question, namely the probability that at least one has it. The numerator describes the probability that exactly one individual has it. Numerical examples are useful here:

Suspect population	frequency	$P(n = 1 \mid n \geq 1)$	$P(S \mid M)$ [Bayes]
10 million	1 in 100 million	0.9	0.9
100 million	1 in 100 million	0.62	0.5
1 billion	1 in 100 million	0	0.09
10 million	1 in 1 billion	1	0.99
100 million	1 in 1 billion	0.9	0.9
1 billion	1 in 1 billion	0.62	0.5
10 million	1 in 10 billion	1	0.999
100 million	1 in 10 billion	1	0.99
1 billion	1 in 10 billion	0.9	0.9

The table contains a comparison with the results yielded by Bayes' rule. The two methods essentially agree. For instance, with a frequency of 1 in 100 million, and a suspect population of 100 million, the chance that the profile is unique is not very high; similarly, the chance that the defendant is the source is only 0.5. Instead, with a frequency of 1 in 10 billion and a suspect population of 1 billion people, the chance of uniqueness is 0.9; similarly, the chance that the suspect is the source is also 0.9. The two methods agree, but they are also different because, unlike Bayes' rule, the uniqueness method cannot easily take into consideration a number of complications, such as the impact of laboratory error rates.

Random match and likelihood ratios. Another popular method to present DNA evidence is by means of the so-called *random match probability*, which expresses the chance that a random individual, unrelated to the defendant, would have the DNA profile in question. This probability is essentially equivalent to the frequency of the DNA profile and it conveys useful information on the probative value of DNA evidence. The problem with this approach is that the random match probability can be a source of confusion. Fact-finders often reason as follows: (i) If the random match probability is, say, 1 in 100 million, then the chance that another individual, not the defendant, has the DNA profile is 1 in 100 million; (ii) this is a very low chance, so the chance that another individual, not the defendant, left the DNA traces must also be very low; (iii) thus, the chance that the defendant is the source of the traces must be very high and it is the inverse of 1 in 100 million. This reasoning is fallacious; it is known as the *prosecutor's fallacy*, which consists in the equation $1 - f = P(S|M)$ (Thompson and Shumann, 1987; Macedo, 2008). The fallacy, which occurs in step (ii), is an instance of the inversion fallacy: the fact-finder has confused $P(M|\neg S)$ with $P(\neg S|M)$ and has assumed that they both equal f. Then, step (iii) follows, because if $P(M|\neg S) = P(\neg S|M) = f$, then $1 - f = P(S|M)$ by the negation rule. Using Bayes' rule avoids the possibility of committing this mistake and it allows for a more complete analysis.

Instead of the random match probability, forensic experts often testify as to the likelihood ratio of DNA evidence, namely $\frac{P(M|S)}{P(M|\neg S)}$.[14] Bayes' rule includes the likelihood ratio as well as the prior odds. So, Bayes' rule is a more comprehensive way to present DNA evidence. Why present only likelihood ratios, then? A couple of reasons bear mentioning. First, we've seen that estimating prior probabilities is not easy, and one might overcome the problem altogether by focusing on likelihoods only. Second, in common law countries, such as the United States and the UK, Bayes' rule might be considered an invasion of the providence of the jury: the jury is not supposed to be told how to reason or how to weigh the evidence. Bayes' rule is precisely a method to weigh the evidence. On the other hand, I think that Bayes' rule is very helpful. Fact-finders are interested in knowing the probability that the defendant is the source, and only Bayes' rule can guide them properly. It might be an invasion of the province of the jury, though suggesting a correct way to reason and weighing the evidence is an essential ingredient to protect defendants and improve trial's accuracy.

5.6 ACCIDENT OR PURPOSEFUL CONDUCT?

In a number of criminal cases statistical evidence is introduced to establish whether a series of events happened accidentally or as the result of purposeful conduct. Suppose a house is on fire four times in a year, and each time the owner is compensated by his insurance company (Feinberg and Kaye, 1991). Did he set fire to his house intentionally?

[14] On likelihood ratios, see (Royall, 1997).

Or suppose that two babies, one after the other, die at the presence of their mother, as happened in the Sally Clark case. Did she kill the babies? Or again, suppose that many infants die in a hospital during a nurse's shifts, as happened in the Lucia de Berk case. Did Lucia kill them? These are cases in which there are two alternative theories: the events happened by chance; events were the result of purposeful conduct. Bayes' rule can help us assess the statistical evidence and draw informed conclusions.

Let us consider the example of the two babies who both died in the presence of their mother. Suppose—contrary to what happened in the Sally Clark case—that reliable statistics show that the chance that two babies, in the same family, die of natural causes is 1 in 100 million. If we reason informally, we might be impressed by such a low probability and conclude that the chance that the mother killed them is overwhelmingly high, and conversely, the chance that they died of natural causes is nearly impossible. This is an example of an intuitive yet bad inference from statistical data. Bayes' rule, instead, offers a principled way to bring clarity to the question and weigh the evidence. Recall the formulation in terms of odds ratio:

$$\frac{P(H|E)}{P(\neg H|E)} = \frac{P(E|H)}{P(E|\neg H)} \times \frac{P(H)}{P(\neg H)}.$$

The two hypothesis that we may want to contrast are that the babies died by natural causes—call it *natural*—and that their mother intentionally killed them—call it *kill*. Now, what is our evidence for the hypothesis *kill*? Our evidence is that both babies died, and from that evidence we want to know how likely it is that the mother killed them. We need to estimate the value of $P(\textit{kill}|\textit{death})$, and Bayes' rule in our case reads:

$$\frac{P(\textit{kill}|\textit{death})}{P(\textit{natural}|\textit{death})} = \frac{P(\textit{death}|\textit{kill})}{P(\textit{death}|\textit{natural})} \times \frac{P(\textit{kill})}{P(\textit{natural})}.$$

The likelihood ratio equals one. If the babies died of natural causes or because they were killed, they would be found dead with equal probability. What really makes a difference are the prior odds. We know that the probability that the babies died of natural causes is 1 in 100 million, so $P(\textit{natural})$ =1 in 100 million. What is the prior probability that a mother kills both her babies? Such a statistic is needed to make a proper assessment here. To have a rough idea, suppose that in a mid-size country like the UK or Germany 1 million babies are born every year of whom 100 are murdered by their mothers. So the chance that a mother kills one baby in a year is 1 in 10,000.

What is the chance that the same mother killed two babies? This is a bit complicated to estimate. We can appeal to a controversial assumption: independence. So, the chance that a mother kills two babies equals 10^{-8}, i.e. 1 in 100,000,000. If we deny independence, we would get a larger probability, maybe 1 in 50,000 or something. Now, assuming independence, both $P(\textit{natural})$ and $P(\textit{kill})$ would equal 100,000,000. This means that the prior odds would equal 1. But with a likelihood ratio of one, the posterior odds equal one, and the posterior probability is therefore 0.5, a value clearly insufficient for a conviction. The lesson here is that, to have a posterior probability greater than 0.5, the prior probability of *natural* must be significantly lower than the prior probability of *kill*. With a fixed likelihood ratio of one, we have:

$P(\textit{kill})$	$P(\textit{natural})$	Posterior Odds	$P(\textit{kill}\vert\textit{death})$
1 in 100 million	1 in 100 million	1	0.5
1 in 100 million	1 in 1 billion	10	0.9
1 in 100 million	1 in 10 billion	100	0.99

As expected, the prior odds play a crucial role in determining whether the mother killed her babies or not. And further, a low probability $P(\textit{natural})$ is meaningless if it is not compared with $P(\textit{kill})$.

5.7 QUANTITY ESTIMATION

To be convicted of a crime such as drug trafficking or grand theft, the prosecutor has to determine the quantity of the drug illegally transported or the amount of money stolen. Often direct evidence of the overall quantity or amount is not available, and thus statistical methods can be used to yield indirect estimates. An example is the *Shonubi* case. To keep things simple, suppose the statistical data establish that drug smugglers like Shonubi carry at least 200 grams on each drug-smuggling trip; Shonubi is known to have made five trips in total although he was found carrying drugs only once. We want to know how likely it is that Shonubi carried drugs on his other four trips. Once again, Bayes' rule can be helpful to fix ideas.

The two competing hypotheses are that Shonubi carried drugs on one trip only, when he was caught, or that he carried drugs on the other four trips as well. Our hypotheses can be abbreviated as follows: '*only once*' and '*five times*.' Besides the statistics, the available evidence is that Shonubi was found carrying 200 grams on one trip, abbreviated '*caught once*.' So, Bayes' rule reads as follows:

$$\frac{P(\textit{five times}|\textit{caught once})}{P(\textit{only once}|\textit{caught once})} = \frac{P(\textit{caught once}|\textit{five times})}{P(\textit{caught once}|\textit{only once})} \times \frac{P(\textit{five times})}{P(\textit{only once})}.$$

First, consider the likelihood ratio. The probability that Shonubi was caught once given that he carried drugs once seems lower than the probability of being caught once while carrying drugs five times in total. On the other hand, the probability of being found with drugs depends on the reliability of the immigration screening procedures. If they are very reliable, e.g. $P(\textit{caught once}|\textit{only once}) = 0.8$, it would be unlikely that over five times, Shonubi is caught only once. We therefore need data about the success rate of custom screening procedures, though it is difficult to obtain such data, because how can we know how successful the screenings are except through the screening procedures themselves? By using the formula for the binomial distribution, here are some numerical examples:[15]

$P(\textit{caught once}|\textit{only once})$	$P(\textit{caught once}|\textit{five times})$	Likelihood Ratio
0.2	0.4	2
0.5	0.1	1/5
0.8	0.001	1/800

Next, let us consider the prior odds. The statistics about other drug smugglers are useful here. They tell us that, with a probability of 0.99, smugglers like Shonubi typi-

[15] We want to know how likely it is that Shonubi would be found carrying drugs only once, given that the probability of being caught each time equals a certain value, such as 0.2, 0.5, or 0.8. Using the binomial distribution seems to be appropriate here.

cally carry 200 grams of drug on each trip. Thus, the prior odds will be 0.99/0.01=99. Putting everything together, we have:

Prior odds	Likelihood Ratio	Posterior odds	$P(\textit{fives times}\mid\textit{caught once})$
99	2	198	0.99
99	1/5	20	0.95
99	1/800	0.1	0.09

Even in the case of estimating quantities, Bayes' rule has proven to be powerful and illuminating. The statistics about drug smugglers who share a certain profile with Shonubi are useful. But in order to properly estimate the posterior probability that Shonubi carried drugs during his five trips and not during one trip only, we need an estimate of the reliability of the custom screening procedure. For the posterior probability will vary dramatically depending on such estimate. It is through Bayes' rule that we can appreciate this and place the statistics in the appropriate context.

5.8 And the guilt probability?

I've illustrated how Bayes' rule can be of help in estimating the probability of some intermediate propositions. Rather than offering a precise way to estimate probabilities, Bayes' rule brings clarity to the factual questions and reveals potential sources of uncertainty. In the Shonubi case, Bayes' rule revealed that we needed an estimate of the reliability of the custom's screening procedures. In the Sally Clark case, it reveled that we needed an estimate of how often mothers kill their children. In the case of DNA evidence, it allowed us to organize, in one elegant formula, different sources of uncertainty, such as laboratory error rates and the possibility of a coincidental match.

The legal probabilists are right when they say that probabilistic tools can improve tremendously the accuracy of the legal fact-finding process, as illustrated in this chapter. This is, however, very different from saying that a high guilt probability is enough to convict. The accuracy-enhancing function of Bayes' rule and of probabilistic tools can be separated from the decision-theoretic rule that a conviction is acceptable when it meets a certain threshold guilt probability. As shown in chapter 2, an appropriately high probability of guilt is not enough to justify a criminal conviction. The claim that a high probability of guilt, in and of itself, is not enough to justify a conviction will be further defended in chapters 6 and 7 that follow.

CHAPTER 6

THE BURDENS OF STABLE JUDGMENT

In some contexts, the high probability of a proposition, in and of itself, does not seem a good enough ground for us to conclude that the proposition is true. Consider this scenario:

> Mark is holding a lottery ticket. Statistically speaking, it is overwhelmingly likely that Mark is holding a losing ticket. If the lottery is fair and only one out of a million tickets is the winning ticket, Mark's chances of winning are one in one million. Yet, Mark does not conclude that he is a holding a losing ticket. He suspends his judgment until he hears (via the newspaper, the TV, the internet, etc.) of the number of the winning ticket.

What is the rationale, if any, behind Mark's behaviour? To begin with, one could say that Mark is too gullible; that he should realize he is never going to win the lottery; that he should stop dreaming. On this view, Mark's behaviour would not be justified at all. I shall put this revisionary position aside. Without being revisionary, here, I am primarily interested in understanding whether we can make sense of, and justify, Mark's behaviour.

If we look at the non-revisionary accounts in the epistemological literature, two features are immediately apparent. First, they aim to provide a rationale for why Mark does not know that he lost (or that he will lose). Second, they are premised on the idea that Mark's lack of knowledge has to do with Mark's evidence. This is plausible. If Mark fails to know, it must be because he fails to have an adequate epistemic justification, and the latter must have to do with Mark's evidence. Unsurprisingly, then, many non-revisionary accounts strive to articulate what it is about Mark's evidence in the lottery scenario that makes it deficient.[1]

Here are a few representative examples of this strategy. Some accounts suggest that Mark's evidence is deficient because it is merely statistical and it fails to be causally or explanatorily connected with the event of winning or losing the lottery (Harman, 1968; Nelkin, 2000). Other accounts suggest that if Mark were to form the belief that he lost the lottery, such a belief would fail to be 'safe' (Williamson, 2000; Pritchard, 2005) or it would fail to be 'tracking' (Nozick, 1981; DeRose, 1996; Roush, 2006). Still other accounts suggest that the evidence is deficient because it does not give Mark enough information to distinguish the winning ticket from the losing ones; the evidence available to Mark only says that every ticket is overwhelmingly likely to lose, but it does not single out the winning ticket (Hawthorne, 2004). Finally, other accounts suggest that Mark's evidence fails to rule out all the relevant or reasonable alternatives to the proposition that Mark's ticket is a loser (Dretske, 1971; Lewis, 1996; Lawlor, 2013).

The proposal I defend here differs from the non-revisionary accounts in the epistemological literature in two respects. First, I am primarily concerned with whether Mark is *justified in suspending judgment*, and not with whether Mark *fails to know*, although the two are related. My proposal is also different from existing accounts in another respect. As noted earlier, most non-revisionary accounts are premised on the idea that there must be something deficient about Mark's evidence, and that such a deficiency is

[1] The exception is Douven (2002). He does not focus on the available evidence, but on the broader inferential consequences. He argues that concluding that a proposition is true on the sole basis of its high probability often leads to *inconsistencies*. The basic idea is that, on the basis of the statistics, Mark would have to conclude that every ticket lost, which would contradict the fact that at least one ticket won. This approach was later refuted by Douven and Williamson (2006).

responsible for Mark's failure to know. Rather than focusing solely on Mark's evidence, I think we benefit from focusing on Mark's epistemic expectations, and in particular, on the following question: *What evidence does Mark expect to acquire at a later time in addition to the statistical evidence he currently has available?* My suggestion is that we should not only look at Mark's current evidence, but also at Mark's expectations to acquire further evidence later on. These expectations, and not Mark's current evidence alone, might offer a better justification for why Mark suspends judgment. The act of suspending (or reaching) judgment should be influenced, I suggest, not only by the strength of the evidence currently available, but also by the expected forthcoming evidence that may undermine or strengthen the currently available evidence. This, at least, will be my guiding idea.

In this chapter, I will outline some principles for when we should suspend or delay judgment and when, instead, we should reach judgment. I will do so by considering some idealized scenarios that are familiar to epistemologists. In the next chapter, I will explore one avenue in which my discussion can be applied, namely to the question of whether a conviction in a criminal trial can rest on a sufficiently high probability of guilt, or whether something more than high probability is required.

6.1 JUDGING, SUSPENDING, AND TENTATIVELY ASSUMING

Let me begin by giving a rough characterization of what I mean by 'suspending judgment' and by 'reaching judgment.' A subject S *suspends judgment about* p when she keeps an open mind about p; when she does not dismiss the possibility of not-p occurring; when she does not settle on p being the case; or when she does not close the question about p. In contrast, a subject S *judges that* p when she settles on p and thus dismisses the possibility that not-p. When S judges that p, in other words, S closes the question regarding p and forms a settled opinion that p is true. A subject S could judge that p at will, without having any ground for doing so. If so, S would be irrational. Instead, I am interested in cases in which S is not irrational, or has some basis for judging that p.

Some might wonder whether the dichotomy between reaching judgment and suspending judgment is too coarse. I think it is. There are plenty of cases in our ordinary life in which we *tentatively assume* that a certain proposition holds, though we haven't properly judged, in a cognitively explicit way, that it does hold. These are cases in which we are neither suspending nor reaching judgment, but rather "tentatively judging" or better—*tentatively assuming* something to be the case. The reasons why we make tentative assumptions, despite the lack of strong supporting evidence, can be different: they can be practical, moral, political, tactical, etc. For instance, when we turn on the car engine to drive home, we typically assume that the engine will work, though we haven't done a thorough examination of the car's conditions. Why do we do that? Possibly, it is because we think that if our car breaks down, we should be able to repair it along the way, so that worrying beforehand isn't worth our while. Interestingly, if we were to begin a long journey—one in which we foresee it will be difficult to find help along the way—we might prefer to have our car thoroughly checked beforehand.[2]

We can helpfully distinguish two levels, which I shall call the *background* and the *foreground*. The background consists of the propositions we take for granted, tentatively assume, or bracket from our epistemic focus. The foreground consists in those propositions whose truth values we are interested in determining. The dynamics of reaching judgment and suspending judgment takes place in the foreground. We reach and suspend judgment relative to those propositions that fall within our focus of current interest. But our focus of current interest cannot comprise everything. Some propositions will lie in the background, either because we take them to be of no interest at all, or because we are tentatively assuming that they are true (or false) for the purpose of our inquiry.

[2] The role of tentative assumptions is even more apparent in policy decision making. For instance, Lindblom (1959) suggested that when administrators make a decision, they do not consider all possible variables, all possible alternative courses of actions and all possible consequences. This comprehensive evaluations of all possibilities would be ideal, yet expecting administrators to do so would be unrealistic. Rather, Lindblom thinks that what administrators can do is simply to "muddle through." This mean to focus on a limited set of variables, compare a limited set of possible solutions, and evaluate a limited set of possible outcomes. As they muddle through, administrators tentatively assume that other possible solutions and other possible variables that could be taken into consideration are not relevant. Practical and tactical considerations tell them to do so, for taking everything into account would mean to paralyze the process of decision making.

Although I will be primarily interested in what happens in the foreground, it is important to keep in mind what lies in the background, as well.

The dynamics of suspending and reaching judgment happens in the foreground: my objective, then, is to formulate some suitable normative principles that govern this dynamics. These principles will identify the conditions under which a subject is entitled to reach or suspend judgment. These conditions can be divided into evidential, practical, and moral. While these distinctions are useful at the analytical level, the principles I offer about suspending and reaching judgment have the peculiarity of blending together evidential, practical, and moral considerations. I think this is a good feature of them, rather than a shortcoming. I shall say why I think so in due course. Still, as the reader encounters these principles and senses a confusion between evidential, practical, and moral considerations, she should be aware that the confusion is intentional.

6.2 JUDGMENT INERTIA

The first principle concerns when one should persist in holding a judgment that she has previously formed. The idea is that, if nothing relevant changes, one should persist in holding a previously formed judgment. Here is an attempt at spelling this out more extensively:

> JUDGMENT INERTIA. Suppose that, at some point t in time, subject S has justifiably formed the settled opinion that proposition p is true. If, between time t and a later time t', nothing changes in S's epistemic and non-epistemic situation, and if p concerns matters that do not suddenly go out of existence (or if S does not expect p to become false before time t'), then S should still be in a position to judge that p is true at a later time t'.

The principle of judgment inertia, in other words, says that if nothing changes that is pertinent to the truth of p and if nothing changes that is pertinent to whether S should still judge that p, then S should persist in the judgment that p. It is hard to identify exactly which changes are relevant. Two examples should help. If I justifiably judge that my car is parked in the street, I am still justified—by judgment inertia—in judging that that my car is parked *at a later time*, provided I have not acquired any information to the contrary. In contrast, if I justifiably judge that it is raining outside, I am not justified in holding the judgment that it is raining a few hours later, because the phenomenon of raining is known to stop quite erratically.

The principle requires that no change in S's evidential *and* non-evidential situation take place. What could change S's evidential situation? In the car example, any relevant evidence—e.g. evidence that a car has been stolen in the neighborhood—can be enough to stop the inertia of my judgment that my car is parked in the street. And what about S's non-evidential situation? This is more difficult to explain. A contextualist would say that if the practical stakes in the situation have risen—e.g. my car is needed tomorrow to go to a job interview and I cannot afford to miss it—then I might not be warranted in maintaining my judgment (Cohen, 1986; DeRose, 2002).

The role of non-evidential considerations can be made more vivid, as follows. Suppose your brain is wired to a screen that can detect your judgments and suspensions thereof. Suppose that, between time t and t', a lunatic has come to power, and that he threats to cut your throat if you still judge that your car is parked in the street. In this case, many non-evidential considerations suggest that you should stop judging that your car is parked in the street. Some might say that this is a far-fetched scenario of little interest to what we actually do. Fair enough; I am not really committed to any particular view here. It is enough to have pointed out that evidential as well as non-evidential considerations *might* play a role.

6.3 SUSPENDING JUDGMENT

I now turn to a second principle, which is about when one is justified in suspending judgment. In its full generality, the principle reads:

> SUSPENSION PRINCIPLE. Let C be a proposition of interest to a subject S—a proposition about which S does not yet have a settled opinion. Even when the probability of C given evidence e is high, a subject S need not draw the flat-out conclusion that C, at time t, provided

> (i) S expects to have new evidence e' about C at a later time t' such that $e + e'$ is better than e alone in establishing or disproving C; and
>
> (ii) S expects that suspending judgment about C between time t and a later time t' will not be unduly burdensome.

The suspension principle is written to handle the lottery scenario at the beginning. What the principle suggests—in particular, what clause (i) suggests—is that the diachronic dimension is important for understanding the lottery scenario. It is true that, regardless of whether we locate Mark before or after the lottery drawing, other things being equal, Mark would be equally justified or unjustified in suspending judgment. This is a consequence of the judgment inertia principle. On the other hand, we should appreciate that Mark *will*—or justifiably expects to—acquire new evidence besides the statistical evidence he has currently available, evidence from other sources, such as a newspaper, the TV, or the internet. In particular, Mark's expectation that he will acquire more evidence about the lottery drawing might affect his current decision to suspend judgment. In this sense, the diachronic dimension is important here. It is important to take into account how the availability of evidence changes across time, and how Mark's expectations are a reflection of that. But why would Mark's expectations be important for his *current* suspension of judgment? Answer: Mark can reasonably expect to hear about the lottery result from a newspaper, the TV, or the internet; then, there is no need for him to rush to the conclusion that he lost beforehand. With some due patience, he will easily acquire firmer evidence about whether he lost or not. And this is where clause (ii) comes into play. Not only is it important that Mark can justifiably expect to acquire more evidence, but also, it is important that his suspension of judgment is not unduly burdensome.

Mark's situation satisfies both conditions in the suspension principle. He expects to acquire new evidence, i.e. evidence from, say, a newspaper, and the new evidence, together with the statistics about his low chances of winning, would constitute better evidence for C. Further, suspending judgment about C is not too burdensome for Mark, because he can simply go on with his life as usual until the newspaper announces the result of the lottery drawing. So, according to the suspension principle, Mark need not conclude that he lost the lottery when he only has statistical evidence available.

I want to address a couple of worries here. First, is it really the case that the newspaper evidence, together with the statistics, will constitute a better body of evidence than the statistics alone? After all, on the statistics, Mark's chances of winning are very low, and Mark should expect that the newspaper will not change things in any significant way; his chances of winning, most likely, will still be very low. In fact, there are two possibilities here. One possibility is that the newspaper announces that Mark's ticket is the winner. This outcome will change things dramatically, and it is what Mark hopes for. The other, more likely possibility is that the newspaper announces that Mark's ticket is not the winner. If so, Mark's chances of losing would increase significantly and they would get closer to certainty. In either case, therefore, the newspaper evidence would bring about a non-negligible change in Mark's probability of losing. More generally, *by using the language of probability* we can say:

> Evidence $e+e'$ is better than evidence e relative to a proposition C whenever the probability of C given e would change significantly—either upwards or downwards—after the addition of e' to the available body of evidence.

The second worry is that we can always expect to amass a better body of evidence in the future, so doesn't it follow that, according to the suspension principle, no one will ever need to draw any conclusion on the basis of currently available evidence? Let me make this worry more vivid. Suppose that, at a later time t', Mark reads the number of the winning ticket on a newspaper, and suppose that, unfortunately, the number is not his. Presumably, Mark would conclude that he lost, and many of us in his situation would do the same. But wait! Mark can expect to acquire more evidence. He can buy another newspaper; he can check the internet; he can watch the news on TV. The newspaper could still be wrong; he could still be the winner! If the suspension maxim holds, it seems, Mark need not draw the conclusion that he lost, even when he reads the newspaper. But then, will he ever be in a position to draw the conclusion that he lost? This is the worry.

My answer to this worry will not be straightforward, but I have a few things to say here. Preliminarily, I want to say that there is no problem if Mark is a bit cautious and does not believe outright what the newspaper tells him. Mark might want to be extra sure and check multiple sources. Mark, however, will have to stop at some point; he cannot go on forever. The problem is that the suspension principle would seem to suggest that Mark need not draw the conclusion that he lost and can keep collecting new evidence forever. This would be a counterintuitive result, suggesting that the principle should be amended or altogether dismissed. I do not think, however, that the alleged counterintuitive result follows, or at least, it does not follow from a proper interpretation of the principle. Let me explain.

The suspension principle justifies one in suspending judgment about a conclusion C on the basis of evidence e if two conditions obtain, i.e. clauses (i) and (ii). If the situation is one in which Mark learns from one newspaper (or more than one) that his ticket is a loser, I think that at least one of these conditions fails. Let's see each condition in turn. Once he reads one or more newspapers, Mark cannot expect to acquire new evidence. Maybe he can look at another newspaper; and then another; and still another; and so on. It does not seem to make much of a difference; this woud be a repetition of the same evidence. Mark might want to consult a few different sources, but he has to stop at some point; he cannot go on forever. At some point, the evidence Mark can acquire will be more or less the same evidence; Mark will simply accumulate more of the same evidence without having any better evidence. In this sense, condition (i) would fail.

Some might be unwilling to concede that condition (i) really fails here. For the sake of argument, let me concede that. Notice that another reason for Mark *not* to suspend judgment comes from condition (ii) which requires that Mark's suspension of judgment should not be unduly burdensome. A clarification on what I mean by 'burdensome' is in order. Simply put, suspending judgment is burdensome if the benefits (utility) are less than the burdens (disutility), under some metric of what is beneficial (useful). I am going to assume that subject S is the one who bears the burden, but this need not be. In fact, others might be affected by my suspension of judgment, so the subject who does the judging need not coincide with the subject who bears the burden. I will, however, identify the two for the time being. We can now see why condition (ii) fails whenever Mark learns that he lost by reading a newspaper (or several newspapers). Suspending judgment would mean for Mark to keep looking for new evidence about the lottery drawing. This behaviour, in the long run, is quite burdensome for him. In contrast, suspending judgment when one has not yet read the newspaper is not so burdensome; it only involves waiting for a few months while going about one's own business in life as usual; it does not involve actively looking for possibly new evidence.

Before moving on, let me illustrate the suspension principle with another, well-known scenario from Goldman (1976);

> Mark is visiting a small village where there are one thousand barns. They are all real except one, which is fake, and the locals inform Mark that the fake barn looks exactly like a real one. Mark happens to stand in front of a barn and it looks real. Statistically, the probability is high that the barn he is looking at is real, yet Mark is unwilling to conclude that the barn is real.

It is instructive to distinguish two sub-cases. Allegedly, if the locals told Mark that one barn is fake, they should know how to distinguish a fake barn from a real one. Before jumping to the conclusion that the barn in front of him is real, Mark can consult the locals and acquire further evidence from them. This would be particularly appropriate if, say, Mark is planning to buy a barn and move to the village. Gathering additional evidence seems possible, and it will not be too burdensome. It might be burdensome in some way, but the burden would be well justified if Mark is planning to buy a barn and he wants to make sure it is not a fake one. In this case, Mark has a good reason to suspend judgment and not conclude that he is looking at a real barn, despite the high probability that the barn is, in fact, real.

The other sub-case I have in mind is more like a skeptical scenario. Suppose the locals tell Mark that there might be a fake barn, which looks exactly like a real one. They cannot identify it, but they have lived happily for many years anyway. Now, this seems a case in which the negation of clauses (i) and (ii) obtains. Mark has no hope to acquire new evidence about the barn; all the evidence he could acquire is unlikely to change anything. This is like a skeptical scenario, in which our senses might be systemically mistaken, yet we have no way to discover it, and thus we end up trusting them one way or another. Does Mark now have a reason to suspend judgment? Now, if Mark is planning to buy a barn; if the locals told him that they had always lived just fine; then, Mark might just as well take a deep breath, conclude that the barn is real, and if he likes it, buy it!

6.4 The evidential and the non-evidential

The burdensomeness condition (ii) does quite a bit of explanatory work here. In both the lottery and the fake barn case, the burdensomeness of having to keep looking for new and better evidence (which is almost surely not available) counsels Mark against suspending judgment. The idea is that a normal individual such as Mark would be unable to easily conduct his life by suspending judgment when he has no hope to gather any better evidence. Yet, a psychologically eccentric individual, who were to follow my suspension principle, will be justified in suspending judgment in many cases which, by ordinary standards, would *not* call for suspending judgment. This individual, presumably, would not care about goods and property, having a place to live, having a settled and pleasant life, or what not. If so, the suspension principle would give him an unlimited license to suspend judgment.

I do not think it is a problem that my suspension principle gives certain individuals an unlimited license to suspend judgment because of their peculiar psychology. The Ancient skeptics, after all, attempted to suspend judgment about their ordinary beliefs and this led them to conduct lives which, by ordinary standards, are unimaginable and impractical. This suggests that the practise of suspending (or reaching) judgment cannot be understood in a mere evidential way. Other considerations, often practical or even moral ones, affect one's justification for suspending or reaching judgment. Another thing to say here is that, although so far I have identified those who do the judging with those who bear the burden, this identification is by no means necessary. Some of the things we judge or refrain from judging to be the case do affect other people. This is very apparent in criminal trials. When jurors decide—judge—that the defendant is guilt, this imposes a tremendous burden on the defendant and only derivatively on the jurors themselves. So, if an Ancient skeptic were to serve as a juror in a criminal trial, it is by no means clear that that he will be justified in suspending judgment about the guilt of the accused no matter what. But on this topic, I shall say more in due time.

All this suggests that our practise of reaching and suspending judgment cannot be merely evidentially driven. It is a fact that in many cases we have only fragmentary and limited evidence. On the basis of this limited evidence, we are often expected to make a judgment. We hardly have full, complete, and unassailable evidence for our conclusions. In light of this widespread situation, we often face a choice between suspending or reaching judgment, e.g. jurors are asked to acquit or convict; legislators are expected to vote in favour or against a certain bill; etc. In all these cases, decisions are not made on the basis of evidence alone. They are made on the basis of evidence, *but also* on the basis of whether more evidence could be gathered in the future if at all, and on the basis of how burdensome it would be to suspend or to reach judgment. These three aspects—i.e. presence of evidence; possible future evidence; burdensomeness—are closely connected.

Some might wonder what would happen if reaching a judgment would be a matter of pure theoretical curiosity, completely disconnected from what is at stake in our lives. I think this would be a case in which we only cared about truth maximization and error minimization. The former means to judge truly or correctly in the maximum number of cases; the latter means to suspend judgment truly or correctly in the maximun number of cases. But even in this idealized scenario, we will care for two goals: truth maximization *and* error minimization. One-sidedly caring for one and not the other goal would mean that we could simply judge everything to be true or that we could always suspend judgment. The two goals, then, must be balanced against one another. But if so, we would have to decide whether we care more about truth maximination or error minimization. And it is impossible to make such a decision on purely epistemic grounds. Hence, evidence alone does not offer sufficient guidance on when we should reach judgment or suspend judgment; evidential and non-evidential considerations must go hand-in-hand.

6.5 Reaching judgment

Let us return to our character, Mark, in the lottery scenario. Some will object to the suspension principle as follows. Suppose Mark has no hope to collect any further evidence besides the mere statistical evidence that his chances of winning are very low. If so, clause (i) of the suspension principle would not apply anymore, yet Mark, in all likelihood, will *not* conclude that he lost the lottery. This shows—so the objection continues—that Mark's unwillingness to draw the conclusion that he lost has nothing to do with whether he expects to acquire more evidence in the future. The objection is instructive because it raises the question of what we should do when we have no hope to collect any better evidence in the future. Is high probability in such cases a good enough ground to draw conclusions? The answer to this question, at first blush, seems negative, as the above objection indicates. Even when Mark has no hope to collect any further evidence, it seems, it does not follow that he should conclude that he lost.

Let us look at the matter more closely. If Mark has no hope to collect any further evidence in the lottery scenario, why is that? It might be because the lottery drawing has been cancelled or because Mark has been forbidden to read newspapers. If the lottery drawing has been cancelled, the proposition *Mark's lost the lottery* would cease to be of interest to Mark. If there is no lottery, it makes no sense to worry whether one has lost or not. Here Mark would not even suspend judgment; he would simply stop thinking about the matter altogether. Hence, in this case, the suspension principle would not apply, for recall that the principle is restricted to propositions that are of interest.

But what if the lottery is still in place and Mark has been forbidden to gather any further evidence about the lottery drawing? To make this vivid, let us imagine that, in principle, Mark can gather further evidence, but if he does so, he will be arrested and severely punished. This is a case in which Mark has no hope to gather any further evidence, not because the evidence is in principle unavailable, but because it would be too burdensome for him to collect it. I am inclined to say that, once more, this would be a case in which the proposition *Mark's lost the lottery* would cease to be of interest to Mark. Understandably, Mark might reason as follows: 'Is it really worth risking my life to find out the number of the winning ticket from a newspaper, given that I have most likely lost the lottery anyway? I should drop the issue altogether and forget about it.' Again, this seems to be a case in which the proposition in question is of no interest anymore, and thus the suspension principle does not apply.

Indeed, the variations on the lottery case I described above are far-fetched. A more realistic scenario is this:

> Mark lives in an authoritarian country and he has been forbidden to read certain books which, allegedly, tell the true story of how the country's political establishment came to power by violence, crime, and deceit. Mark wants to know the truth about his country's history, but the price for the truth will most likely be his life.

This is a case in which, depending on how much Mark is devoted to the truth, as it were, the history of his country will, or will not, cease to be an issue of interest to him. Let us suppose Mark is very devoted to truth. Maybe, his dad was assassinated during the coup d'etat which brought to power the current political establishment. Now, Mark wants to find out the truth, but he cannot access the forbidden books. This would seem to be a case in which Mark suspends judgment, but again, the objection arises: Since Mark suspends judgment even if he has no hope to gather further evidence, wouldn't this show that the suspension principle is inadequate? I think not. Mark's impossibility to recover any further evidence is only temporary. He suspends judgment because he has hope that, one day, he will uncover the truth; he has hope that, one day, the current political establishment will be overthrown; etc. Further, Mark believes, and understandably so, that his state of suspension of judgment is not too burdensome for him anyway, for he will not be arrested if he suspends judgment. So, both conditions in the suspension principle are satisfied, and that is why Mark may suspend judgment about the history of his country.

Contrast the above case with this:

> Mark is the advisor to the justice minister, and he has collected statistical evidence about whether harsher sentences have the effect of deterring criminals. On the basis of the statistics, Mark can say that there is a high

> probability that harsher sentences have a deterrence effect. As he shares his findings, the minister presses Mark on whether he has collected enough evidence; on whether it might be better to look at the data more carefully; and on whether it might be better to perform additional studies. The justice minister, however, realizes that he lives in a country with skyrocketing crime rates, in which many other measures have failed and in which any further delay in taking action will endager the safety of citizens. Should the minister conclude that harsher sentences have a deterrence effect? And what about Mark?

I am not sure what Mark should conclude. What is at stake for him seems much less pressing than what is at stake for the justice minister. Understandably, the minister would convince himself that Mark's statistics are good enough and he will take action by proposing appropriate legislation. What this suggests is that the high probability of a conclusion C might become sufficient to draw the conclusion in question when no further evidence is forthcoming between the current moment and some later time, *and in addition*, when suspending judgement until some later time would be too costly. More precisely, here is a (weak) version of the judgment principle:

> JUDGMENT PRINCIPLE (WEAK). Let C be a proposition of interest to a subject S—a proposition about which S does not yet have a settled opinion. When the probability of C given evidence e is high, a subject S should draw the flat-out conclusion that C, at time t, provided
>
> (*) S does not expect to have new (possibly contradicting) evidence e' about C before time t';
>
> (not-ii) S expects that suspending judgment about C between time t and a later time t' will be unduly burdensome.

The judgment principle indicates that the normative reasons to draw a conclusion form a complex structure of evidential, semi-evidential, and non-evidential reasons. The evidential reason is that the probability for the conclusion is sufficiently high; the semi-evidential reason is that no further, possibly contradicting evidence is forthcoming; the non-evidential reason is that persisting in suspending judgment is too harmful for the interested parties.

Let's now see the principle in action. Recall the scenario involving Mark and the justice minister who wants to reduce crime rates. The scenario is one in which contradicting evidence might be forthcoming, yet the minister decides to convince himself that the available evidence is enough for concluding that harsher sentences deter potential criminals. The minister, presumably, thought that between t and t' no further, potentially contradicting evidence would be forthcoming. He knew that, in order to obtain new evidence, new research would need to be conducted, and that this would require some time. So clause (*) of the principle seems satisfied, although more evidence could be available later on after time t'. As for clause (not-ii), the minister was aware that failing to take action as soon as possible would be extremely harmful for his country (and possibly, for his reputation). Thus, if the minister decided to conclude that harsher sentences do have a deterrence effect, he would act in accordance with the judgment principle above.

Some might object here that the minister has not really judged that harsher sentences have a deterrence effect. Rather, he has convinced himself that this is the case for the purpose of taking action. I do not see much difference between the two formulations, and I do not want to enter a terminological dispute. In the previous section, I've insisted that judging does not take place in a vacuum and that reaching and suspending judgment are not a matter of pure theoretical curiosity. Consequently, it should not be surspring that the minister's judgment is not one purely based on evidence; rather, it is tightly intertwined with the actions the minister intends to take after his judgment.

I am willing to concede, however, that the first formulation of the judgment principle might rely too heavily on practical considerations, and too weakly on evidential ones. By way of comparison, then, it is useful to formulate an epistemically stronger version of the principle, as follows:

> JUDGMENT PRINCIPLE (STRONG). Let C be a proposition of interest to a subject S—a proposition about which S does not yet have a settled opinion. When the probability of C given evidence e is high, a subject S should draw the flat-out conclusion that C, at time t, provided
>
> (*) S does not expect to have new evidence e' about C at a later time t'

> *and*
>
> (**) even if e' became available, S does not expect evidence $e + e'$ to be better than e alone in establishing or disproving C; and
>
> (not-ii) S expects that suspending judgment about C between time t and a later time t' will be too burdensome.

The stronger version differs from the weaker version because of condition (**), requiring that the evidence e be immune from improvement or refutation by other evidence.[3] This means that condition (**) requires that the probability of C given $e+e'$ be the same as (or roughly the same as) the probability of C given e. The judgment principle, in its strong formulation, applies to many of our ordinary beliefs, e.g. that we have hands, that the lamp is red, that our house is where we left it this morning, etc. For one thing, we do have good evidence that these beliefs are true, though their probability is still short of one because our evidence is fallible. For another thing, we cannot hope to gain any better evidence—e.g. we cannot gain any better evidence that we have hands besides looking at them—and suspending judgment about our ordinary beliefs would put us, or least most of us, in an unbearable condition.

Some might wonder whether a purely epistemic judgment principle, one without clause (not-ii), would be appropriate to certain situations. These would be situations in which time constraints are wholly absent. This might occur when people make extremely thorough, well-researched, and careful judgments, without being subject to any time constraints in their research. In these situations, people strive to say what should be the definitive word on a topic. I think these cases are excessively idealized, for even in academic research there are constraints in time, resources, social relevance, etc. Nevertheless, it is useful to think of a purely evidential judgment principle as applying to limiting cases. This can help us distinguish evidential and non-evidential considerations in the making of judgments and suspensions thereof.

[3] One way to understand this requirement is in terms of "stable" or "resilient" probability (Skyrms, 1977; Leitgeb, 2010).

6.6 WHO BEARS THE BURDEN?

So far I have been speaking loosely of costs and burdens which are born by the interested parties. But who are those parties? The subject who gathers the evidence need not be the same as the subject who bears the burden of judgment. I should make clear that I am interested in the burden *of the consequences of* a judgment. I shall simply speak of the burdens of judgment in the sense that a judgment is associated with some more or less burdensome consequences. This is apparent, for instance, in the legal fact-finding proces. In a criminal trial, the fact finders, i.e. the jurors or a judge, reach the judgment—they convict—but they are only marginally touched by the consequences of their judgment or suspension thereof. The most interested party here is the defendant and society overall; the prosecutor, the defense attorney, the judge, the jurors are interested parties only derivatively. (I will discuss criminal cases extensively in the next chapter). In civil cases, instead, it seems that the burden of a decision is born by the plaintiff and the defendant in possibly uneven ways. To make the point more vivid, consider this scenario:

> Mark and Jim went hunting. At some point they aim at a fox and shoot in its direction. They fail to catch the fox, but unfortunately, they end up injuring a person behind the bushes. The victim has to amputate a leg and wants to recover damages from whoever is responsable for the injury. Crucially enough, it is clear that only one bullet reached the victim's leg, so either Jim or Mark is responsable. The issue is who.

In such a situation, U.S. law is that Mark and Jim are jointly liable unless they can individually establish that they are not responsible for the victim's injury.[4] Typically, the plaintiff has to establishes that Mark or Jim caused the injury, but in this case, the plaintiff is unable to recover evidence that points to Mark or Jim in particular. Importantly, the plaintiff, because of no fault of his own, cannot recover any more evidence about the incident.

What are the available options to decide the case? One option is to suspend judgment by declaring that Mark or Jim cannot be held liable, until new and better evidence is

[4] Summers v. Tice, 33 Cal.2d 80, 199 P.2d 1 (1948). See also (Wright, 2008; Gifford, 2005).

produced. This decision would be tremendously burdensome for the plaintiff who would receive no compensation at all. It would be a situation that violates clause (ii) in the suspension principle. This suggests that suspension of judgment might not be the best option. On the other hand, holding Mark or Jim liable would be unacceptable. If at most one of the them is responsable, and no evidence can say who is, their probability of being responsabile is .5. This cannot be enough for a finding of civil liability. So, holding both liable might not be the best option either. We have reached an impasse.

The suspension principle should draw our attention to at least two issues, closely related to the two conditions in the principle. One issue—which I've already discussed—is, who bears the burden of the consequences of the final decision and what is this burden? The other issue is, how can more evidence become available? It is this second issue that U.S. law focuses upon. By holding Mark and Jim—tentatively—jointly liable unless they can prove their innocence, U.S. law forces Mark and Jim to produce more evidence, on the assumption that they are in a better position to do so. This is a sensible strategy and it is coherent with the suspension principle. Civil cases differ from criminal ones, among other things, because in the latter cases the burden of producing evidence lies almost exclusively on the prosecutor, and thus, when the prosecutor fails to produce adequate evidence, this simply means that the case against the defendant is dismissed.

***

To recapitulate, when we ask whether a judgment, or suspension thereof, is warranted or not, we can profit by taking into consideration the following variables:

(-) How probable is C given e;

(-) Whether there could be more evidence e' about C that could modify the probability of C given e;

(-) Who should bear the burden of presenting more evidence, and what this burden is;

(-) Who bears the burden of the consequences of reaching or suspending judgment, and what this burden is.

6.7 Toward criminal trials

I examined how the (envisioned) acquisition of possible future evidence affects our patterns of judgment. I argued that we are justified in suspending judgment when we expect to acquire better evidence. The situation in criminal trials is somewhat different. The fact-finders are forced to make a difficult decision. For one, they are not allowed to suspend judgment on the ground that new evidence could become available. In fact, they are sometimes explicitly told to disregard certain items of evidence, and they are not allowed to gather any new evidence on their own. For another, they are expected to reach a judgment of guilt only if the highest standard of proof has been met. Given the gravity of consequences of their findings, the fact-finders cannot ignore the possibility of future evidence, but how can this be reconciled with the need to arrive at a final decision? In the next chapter, I attempt to lessen this tension by articulating an account of what the criminal standard requires. My account will take into consideration both the need of finality and the need to satisfy the most stringent standard of proof.

CHAPTER 7

FOUR WAYS A REASONABLE DOUBT CAN ARISE

Ask a judge or a lawyer what *beyond a reasonable doubt* means, and you might hear something like this: it does not mean beyond any possible doubt; it means beyond any plausible, realistic, non-contrived, substantial doubt. These are paraphrases, but do they advance our understanding? The U.S. Supreme Court in Holland v. U.S. (1954) has discouraged any definition of the criminal standard of proof because, as the Court put it, 'attempts to explain the term "reasonable doubt" do not result in making it any clearer' (348 U.S. 121, 140). Dershowitz (1997) has called this 'an act of abject intellectual cowardice' (p. 69) and Laudan (2006) has documented the disagreements among legal practitioners about the meaning of the criminal standard of proof. Some might think this is a scandal: we do not understand what the standard means, but we still send people to jail and even to death!

Some conceptual clarity is certainly called for here. Yet, we should not expect a definition that can tell us—in full generality and with no ambiguity—when a defendant's guilt has been established beyond a reasonable doubt (BARD, for short). This would be too much to expect. We are, after all, talking about a standard, not an algorithm. If legal practitioners cannot articulate what BARD means, this is not *ipso facto* a reason for thinking that the practice of acquitting and convicting people is completely arbitrary.

Against this background, I wish to clarify the meaning of BARD, while being wary of the inherent limitations in doing so. In the existing literature, one prominent view is legal probabilism. It is the view that establishing guilt beyond a reasonable doubt means to establish that the defendant's *probability of guilt*, given the total evidence presented at trial, meets a threshold, say, 0.99 or 0.999 (Kaplan, 1968; Kaye, 1999; Tillers and Gottfried, 2007). This definition is simple, crisp and elegant, but a too literal interpretation of it is obviously problematic. If a probabilistic threshold is understood as a criterion which the fact-finders should mechanically apply whenever they confront the decision to convict or acquit, two difficulties arise. First, it is not clear where, exactly, the threshold should be placed: is it 0.99, 0.89, 0.899, 0.999, or what? A second difficulty is that assigning a probability value to guilt itself might not be feasible. But these difficulties can be sidestepped if we understand legal probabilism less mechanistically. The legal probabilists can defend their proposal by conceding that they are not offering a recipe that should be directly implementable in court. Assigning probabilities to propositions, they could say, is an idealized process, a regulative ideal which can improve trial proceedings. In this spirit, setting a probabilistic criterion for criminal convictions would only be a way to theorize about the meaning and function of the criminal standard of proof.

In response to legal probabilism, many have argued that probability—or at least, probability alone—is of little help in clarifying what the BARD standard requires (Cohen, 1977; Nesson, 1979; Thomson, 1986; Walton, 2002; Stein, 2005; Pardo and Allen, 2008; Ho, 2008; Haack, 2011). A number of different positions are represented here. Some argue for a total dismissal of legal probabilism and others for a refinement of it. I am among those who think that legal probabilism needs refinement. Criminal trials are sophisticated procedures, and many considerations go into the weighing and assessment of the evidence. My main criticism of legal probabilism is this: the criminal standard of proof should take more into account than the probability of guilt on the available evidence.

The account I will articulate is indebted to Allen (2010). His view is that BARD requires that there be no plausible alternative story to the plausible story of guilt. A plausibility-based approach is appealing, but the obvious problem is that the notion of a ‘plausible story’ is left under-defined. In contrast, a probability-based account rests on a fully worked out mathematical theory. Despite its mathematical underpinnings, however, it is hard to relate the notion of probability to actual trial proceedings: jurors do not naturally assign a probability to guilt, and it is difficult to do it even if we wanted to. On the other hand, the notion of plausibility seems closer to how jurors actually reason in trial proceedings. Both accounts have merits: one is theoretically grounded in a mathematical theory and the other is closer to legal practise. We can make progress if we merge the two in a suitable way, and this is what I plan to do here. The account of BARD I plan to propose takes seriously the idea that a high guilt probability is a requirement for a criminal conviction, but this requirement cannot be, in my view, the only one. I will argue that the criminal standard of proof requires, among other things, that the prosecutor offer a reasonably specific narrative, story, or reconstruction of the crime.

7.1 Relevance

Determinations of guilt or innocence must be based on the evidence presented at trial; this much seems clear. But what are we to understand by ‘evidence’? This section offers an account of evidence in criminal trials which will constitue the foundation for my explication of the criminal standard.

In a trial, the prosecution and the defense introduce various items as evidence: assertions made by lay witnesses or expert witnesses during depositions, direct-examination and cross-examination; documents and records; written declarations; exhibits; etc. Why are some items presented as evidence and not others? The principal reason—leaving aside the legal issue of admissibility—is that some items are regarded as relevant evidence. The *Federal Rules of Evidence* (F.R.E. for short) define ‘relevant evidence’ as follows:

> Evidence is relevant if: (a) it has any tendency to make a fact more or less probable than it would be without the evidence; and (b) the fact is of consequence in determining the action. F.R.E., 401.

To fix some terminology, I distinguish between (i) the supporting evidence, often called the *fact probans*; (ii) the intermediate proposition to be proven, often called the *fact probandum*; and (iii) the ultimate fact probandum or the action, which boils down to the issue of *guilt/innocence*. It is also useful to distinguish between the *probative value* of a fact probans for a fact probandum and the *materiality* of a fact probandum for the ultimate fact probandum. With this terminology in hand, we can understand part (a) of rule 401 as referring to the probative value of a fact probans for a fact probandum, and part (b) as referring to the materiality of a fact probandum for the ultimate probandum.

My usage of the word 'fact' might sound odd to philosophers, but hopefully it sounds more natural to lawyers. A fact, as I use the term, is a state of affairs, a such-and-such arrangement of things. I do not take a fact to be *eo ipso* true, but only capable of being true (or false). In accord with legal practice and etymology, a 'factum' is something that is made, constructed, build up, put forward (Shapiro, 2003). The fact probanda and the facta probantes are put forward by the prosecution or the defense. It is the task of the fact-finders to decide whether such proffered facts are true or false.

7.1.1 Generalizations and more

Let us now return to the notion of relevance. According to part (a) of the definition, an item counts as a fact probans provided it increases (or decreases) the probability of a fact probandum. Note that fact probans and fact probandum are relational notions; they make little sense in isolation from one another. As Henry Wigmore writes:

> Evidence is always a relative term. It signifies a relation between two facts, the *factum probandum*, or proposition to be established, and the *factum probans*, or material evidencing proposition . . . No correct and sure comprehension of the nature of any evidentiary question can ever be had unless this double or relative aspect of it is distinctively pictured in each instance

> (Wigmore, 1913, p. 5).[1]

Evidence law scholarship rightly emphasizes that certain facts become evidence for other facts only through the application of regularities, generalizations, background assumptions, common and expert knowledge, causal knowledge, analogies, etc.(Schwartz, 2011; Fisher, 2008; Wigmore, 1913). Let's look at an example:

> *Example 1.* A DNA evidence expert asserts "the defendant's DNA profile matches with the semen on the victim's body." The assertion, presumably, is evidence for the fact that the defendant had intercourse with the victim. Why?

An argument chart—*see* chart on page 186—can make it explicit that generalizations and domain-specific or expert knowledge are needed to infer from the expert testimony (fact probans) the conclusion that the defendant had intercourse with the victim (fact probandum). The chart visually represents a piece of reasoning in which the probantum plays the role of the conclusion and the facts probantes, generalizations, and expert knowledge play the role of the premises.[2] So, we can say that an item counts as evidence for a fact probandum so long as there is a *prima facie* acceptable argument that contains the said item among the premises and that contains the probandum as the conclusion (Schum and Kadane, 1996).[3] Of course, this *prima facie* argument can be rebutted, for typically it is a non-deductive argument, i.e. one in which the truth of the premises does not necessitate the truth of the conclusion, but simply makes it more probable.

[1] Alone similar lines, the Advisory Committee writes in the notes to the F.R.E., relevant evidence 'is not an inherent characteristic of any item of evidence but exists only as a relation between an item of evidence and a matter properly provable in the case;' see also (James, 1941).

[2] (Wigmore, 1913) famously devised a chart method to analyze trial evidence; see also (Twining, 1986; Schum and Kadane, 1996; Goodwin, 2000; Anderson et al., 2005). My chart is a simplification of a Wigmore's chart.

[3] Along similar lines, Ronald Allen (1991) writes: 'Evidence takes on meaning for trials only through the process of being considered by a human being. The "formal" evidence initiates trains of reasoning that then form the basis for deliberation (p. 1103).' In other words, Allen is suggesting that an item of fact cannot qualify as evidence independently of the inferences that can be drawn from it.

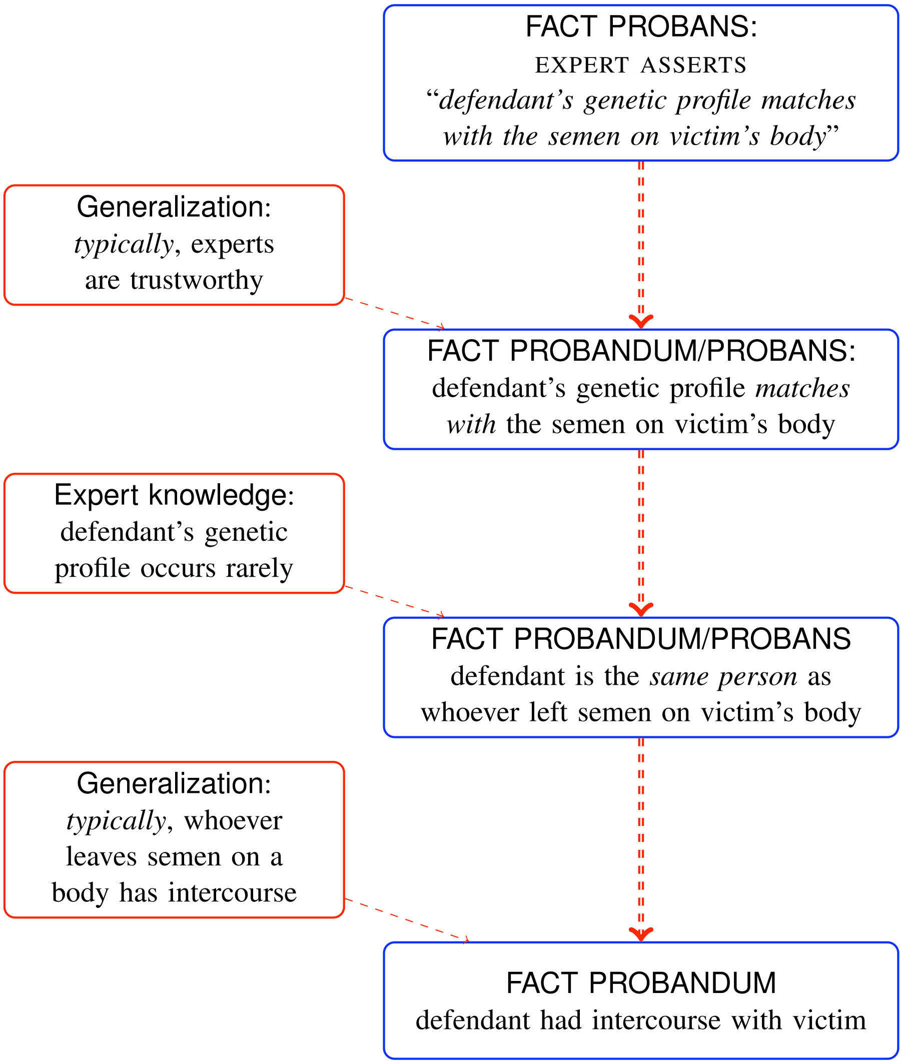

FACT PROBANS:
EXPERT ASSERTS
"*defendant's genetic profile matches with the semen on victim's body*"
Generalization:
typically, experts are trustworthy
FACT PROBANDUM/PROBANS:
defendant's genetic profile *matches with* the semen on victim's body
Expert knowledge:
defendant's genetic profile occurs rarely
FACT PROBANDUM/PROBANS
defendant is the *same person* as whoever left semen on victim's body
Generalization:
typically, whoever leaves semen on a body has intercourse
FACT PROBANDUM
defendant had intercourse with victim

Figure 7.1: The diagram shows that when we draw inferences from the evidence, we rely on generalizations, background and expert knowledge.

7.1.2 Probative value, verbalization, and narratives

The role that generalizations, background assumptions, common and expert knowledge, etc. play in assessing the probative value of an item of evidence is widely recognized. We should take this a step forward. It is often the case that the same fact probans, taken in isolation, can be evidence for contradictory conclusions. Consider this example:

> *Example 2.* A police officer asserts "I found the defendant with a bloody knife in his right hand; he was standing next to the dead body, and the victim had a fresh, deep wound in her chest." The assertion, presumably, is evidence for the fact that the defendant killed the victim. Is that really so?

Our common-sense conception of the world suggests that—plausibly, probably, or more often than not—if one is found with a bloody knife next to a dead body, he must have had something to do with the killing. But we should be careful, for we can imagine two different scenarios here. First, imagine that the defendant was stabbing the victim, and that at the end of this, the police officer saw the defendant with a knife in his hand. On this scenario, the officer's assertion is evidence that the defendant killed the victim. In contrast, imagine that the defendant heard the victim scream, he came to rescue her, and with horror, he found a knife in her chest. Imagine that, as the defendant was pulling out the knife, a policemen arrived and saw the defendant with a bloody knife in his hand. In light of the second recounting of the events, the assertion that the defendant was holding a bloody knife in his hand is anything but evidence that he killed the victim. It is actually evidence for the conclusion that he did his best to help the victim and save her life. The upshot is that any item presented as evidence needs to be placed in a larger narrative within which its probative value, relative to a probandum, can be properly appreciated.

Some might worry that I am going around in a circle here. I am suggesting that the probative value of an item of evidence depends on the larger narrative in which the item can be placed, but—and this is the circularity—if we knew which narrative to select, we would not need the item of evidence in the first place. All I am suggesting, however, is that the police officer's assertion "the defendant was holding a bloody knife in his hand" can be located in diametrically opposite narratives, each consistent with the officer's assertion. We should take note of this fact.

Let me address the circularity worry in a different way. Without going so far as to say that items of evidence are infinitely interpretable, we should nevertheless recognize that trials are predominately verbal and comunicative events. The evidence is presented to the fact-finders through the direct- and cross-examination of witnesses, and even the presentation of physical evidence is often mediated by a verbal description of it. This means that whichever party introduces an item as evidence will communicate and verbalize it for the fact-finders in a particular way, most likely in the way which makes the item fit well with the crime narrative the party in question wants to propose. Along these liens, Robert Burns writes:

> [T]he existence, proper description, and meaning of even the most basic of circumstantial evidence at trial is partly determined by its place in different levels of narrative. They will thus be partially determined, 'colored' if you wish, by whatever renders those narratives more acceptable. (Burns, 2004, p. 169)

Given the inherent need to verbalize and communicate the evidence to the fact-finders, the proposed narrative and the proffered items of evidence determine each other: the items of evidence support the narrative, and conversely, the choice of a particular narrative shapes the verbal presentation of the items of evidence. An hermeneutical circle between the interpretation or verbalization of the evidence and the construction of the narrative seems hard to escape.

7.1.3 Materiality and narratives

The importance of larger narratives is even more apparent if we turn to part (b) of the definition of relevant evidence. Part (b) requires that the fact probandum for which evidence is offered be 'of consequence in determining the action.' In other words, the probandum must bear on, or be material for, the question of guilt or innocence. In the example with the bloody knife, it is obvious why the fact that the defendant killed the victim would bear on the issue of guilt. But things need not be so clear.

Suppose a witness asserts "the defendant was near Union Square in San Francisco at 6:30 PM." With the help of a suitable generalization, this assertion tends to prove that the defendant was, in fact, near Union Square in San Francisco at 6:30 PM. Is this fact material for the issue of guilt or innocence? It depends. In principle, any fact could bear on the issue of guilt or innocence. The deciding factor here should be whether the fact in question can be inserted in a larger narrative which, if taken at face value, tends to prove guilt or tends to weaken the inference toward guilt. To determine materiality and relevance, the narrative need not be explicitly stated or fully worked out; it simply needs to be suggested, by either the prosecution or the defense, so that the materiality of the fact probandum becomes apparent to the fact-finders and the presiding judge.

Let me give an illustration, from the standpoint of the defense, of the importance of locating a fact probandum in a narrative. In a bank robbery case, whether a witness played soccer seems at first quite irrelevant. But now consider this cross-examination of the prosecutor's witness:

> Defense Lawyer: Do you recognize the defendant as the man who robbed the bank?
> Witness: Yes, I do.
> DL: Did you examine some pictures at the police station?
> W: Yes, I did.
> DL: Could you tell us how the photographic identification went?
> W: They showed me a series of pictures and I recognized one.
> DL: I understand. Do you play any sport?
> W: I am sorry – what?
> DL: I asked whether you played any sport.
> W: Why is this relevant?
> DL: Please, answer my question.
> W: Yes, I play soccer regularly on a team.
> DL: I would like to show you a photograph. [*Showing the photograph to the witness*] Do you recognize anyone?
> W: I am in the picture, and so are the people in my team.
> Prosecutor: Objection – this is irrelevant.
> DL: I am sorry. I need to ask a few more questions, your Honor. When was the picture taken?
> W: Last summer.
> DL: On the 4th of September?
> W: I believe so.

> DL: That is, one month before the robbery?
> W: I suppose so.
> DL: Did you know the people on the other team?
> W: Some of them, not everybody.
> DL: Can you please look at the photo and tell us if you can recognize anyone?
> W: I know this one; this one as well; this ...
> DL: Let the record show that the witness recognizes the man in the second row, third position from the right. So, what were you saying?
> W: This one...looks like...He looks like the man in the picture I identified at the police station!
> DL: As a matter of fact, they are the same person.
> (Carofiglio, 2006, p. 56). Translation from Italian is mine.

At some point during the cross-examination, the prosecutor even objected to the defense lawyer's line of questioning. It looked as though the defense lawyer was dwelling on irrelevant facts, which later turned out to be anything but irrelevant. In conducting the cross-examination, the defense lawyer had in mind a narrative of what happened—i.e. roughly, that the witness went to the police station; that he was shown a picture and that he identified a familiar face, not the robber's face. It is only within such a narrative that the fact that the witness played soccer becomes relevant. It becomes relevant because it shows that the witness might have misidentified the defendant as the perpetrator.

The examples could continue, but I shall stop here.[4]

***

Let me recapitulate what I've said so far. First, an item of evidence counts as *probative* evidence for a probandum through background assumptions, generalizations, domain-specific knowledge, etc. Second, the probandum counts as *material* for the issue of guilt or innocence if it can be part of a larger narrative which tends to prove or disprove guilt. This two-pronged characterization mirrors quite closely parts (a) and (b) of rule 401 in

[4] A source to find illustrative examples are cross-examination handbooks; see e.g. the classic (Wellman, 1903) and more recently (Clark et al., 2010).

the F.R.E.

I shall say more on the notion of a narrative in the next section. For now, a narrative is simply a plausible “web of occurrences” in which a fact probandum can be located to support the conclusion that the defendant is guilty or innocent. Often, in order to assess relevance, there is no need to spell out a narrative, not even a sketch of it. If the fact probandum in question is that the defendant killed the victim, there is no need to hint at a narrative that can justify why this fact matters for the issue of guilt. But when the fact probandum concerns the defendant’s whereabouts or a witness’ hobbies, the prosecution and the defense should communicate to the fact-finders the larger narrative they have in mind (or at least, they should offer hints or sketches of it).

7.1.4 Narratives and probability

Some might wonder whether the legal probabilists will accept the characterization of relevant evidence I’ve proposed. Presumably, the probabilists will say that an evidential item E counts as probative evidence for (against) a probandum p whenever the probability of p given E is higher (lower) than the probability of p without E.[5] The probabilists, however, have no reason to deny that generalizations, background assumptions, domain-specific knowledge, etc. play an important role in determining whether the probability of p given E is higher or lower. Probabilities, after all, are not estimated in an informational vacuum.

The probabilists might have a harder time accepting the predominant role that narratives play in my account of relevance. The probabilists, however, have two ways to incorporate narratives in their account. First, they can follow closely the definition of relevant evidence in F.R.E., rule 401. They can easily read part (a) probabilistically, e.g. as the requirement that, for an item to count as probative evidence for p, it must increase or decrease the probability of p. Next, they can read part (b) as putting some constraints on what p can be like. As rule 401 suggests, proposition p must be of consequence to determining the action, or in my terminology, p must be part of a larger narrative which

[5] Or slightly differently, the probabilists think of probative value in terms of likelihood ratios, i.e. E is probative of p whenever $\frac{P(E|p)}{P(E|\neg p)} > 1$.

tends to prove or tends to disprove guilt. This is a somewhat indirect way to incorporate narratives in a probability-based account of relevance. But the probabilists can use narratives more directly. They can think of relevant evidence as evidence that increases the probability of a narrative which the prosecution or the defense are proposing. Instead of looking at whether an evidential item increases the probability of isolated propositions (which, in turn, are material for guilt or innocence), the probabilists can directly consider the probability of an entire narrative. The switch from isolated propositions to narratives is not in contradiction with a probability-based account of relevant evidence. Both isolated propositions and entire narratives, after all, can be more or less probable on the evidence.[6]

7.2 NARRATIVITY

Given the importance of crime narratives in my account, I will look more closely at their role in legal fact-finding, their construction and general structure. I will conclude this section by offerring a first requirement as part of my explication of the BARD standard; I call it the *narrativity* requirement.

7.2.1 Guilt simpliciter or a narrative of guilt?

Some might hold that at trial the prosecutor should establish guilt, not a narrative of guilt. The contrast here is between those who think that establishing guilt simply means establishing a number of propositions which jointly establish guilt (*atomistic model*), and those who proceed more holistically by offerring an incriminating narrative, a cohesive reconstruction of what happened (*holistic model*). I think it is difficult to establish

[6] The second, more direct way to incorporate narratives in a probability-based account of relevance has the virtue—or maybe the vice—of blurring the distinction between materiality and probative value. Traditionally, an item counts as relevant provided it is probative of a proposition that is material for guilt or innocence. On my proposed characterization, instead, an item counts as relevant if it is probative of a crime narrative. The probative value of an item of evidence, relative to a proposed narrative, can be understood probabilistically—i.e. the item makes the narrative more probable—or non-probabilistically—i.e. the item evidentially supports the narrative (in a sense to be specified).

guilt without offerring a narrative of guilt, especially because a crime involves the concomitant occurrence of the *actus reus* and the *mens rea* (Fletcher, 1998; Kaplan et al., 2008; Lippman, 2010). Spelling out the tight connection between the two is hard to do when trying to establish isolated propositions. To illustrate, consider an abridged definition of first-degree murder from the *California Penal Code*:

> Sec. 187. Murder is the unlawful killing of a human being, or a fetus, with malice aforethought. [. . .]
>
> Sec. 188. Such malice may be express or implied. It is express when there is manifested a deliberate intention unlawfully to take away the life of a fellow creature. [. . .]
>
> Sec. 189. All murder which is perpetrated by means of a destructive device or explosive, a weapon of mass destruction, knowing use of ammunition designed primarily to penetrate metal or armor, poison . . . [*long list of other possibilities*] . . . is murder of the first degree.

Let us suppose that a prosecutor, with the intent of establishing that the defendant is guilty of first-degree murder, sets himself to establish that (1) the defendant killed another human being; that (2) the defendant did so with express malice; and that (3) the defendant killed the victim through poisoning. It is hard to imagine how the prosecutor could prove these three propositions in isolation from one another, without offerring a well-specified, unifying narrative of what happened. At best, he might be able to prove (1) and (3) without offerring a narrative. For instance, suppose that expert medical testimony shows that the victim died of poisoning and that a search recovered a small quantity of poison in the defendant's house, the same type of poison which caused the victim's death. Possibly, these two items of evidence could be enough—absent any unifying narrative—to establish propositions (1) and (3). But what about proposition (2)?

The California Penal Code is clear that, in order to establish express malice, the prosecutor must show that the defendant 'manifested a deliberate intention unlawfully to take away the life of a fellow creature.' It seems difficult to establish (2) without offerring a more well-specified narrative of what happened before and after the killing.

What if the defendant administered the poison by accident? What if the defendant actually wanted to kill himself and not the victim? What if the poison was placed in the defendant's house by a third party? These questions, which all pertain to proving or disproving proposition (2), can only be answered by offerring a narrative of what happened before, during, and after the killing.

Let us look at another example. The *Model Penal Code* defines a weaker *mens rea* requirement, i.e. negligence, as follows:

> A person acts negligently with respect to a material element of an offense when he should be aware of a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that the actor's failure to perceive it, considering the nature and purpose of his conduct and the circumstances known to him, involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation. Sec. 2.02 (2) (a).

Note the intermixing of factual language (e.g. 'a person acts') with normative language (e.g. 'unjustifiable risk'). The definition also openly requires the prosecutor to offer an account of the defendant's conduct which takes into consideration its 'nature and purpose' and whether it amounts to 'a gross deviation' from a reasonable person's standard of care. This requires the prosecutor to offer a quite detailed narrative of what happened—a narrative that takes into consideration the *totality of the circumstances*. (More modestly, the prosecutor can also offer a set of alternative narratives of how things supposedly went. He can admit that this is the best he can do and that he is unable to select one specific narrative of the facts. This indecisiveness might weaken the prosecutor's case, but not significantly, provided that, on each of the proposed narratives, guilt can be established.)

In what follows, then, I will assume that the prosecutor is expected to prove a narrative of guilt, not guilt simpliciter, or that establishing guilt just means establishing a narrative of guilt. Also, in speaking of a narrative, I will mostly refer to the prosecutor's incriminating narrative, because the defendant is not expected to offer a full-fledged narrative of the crime; the burden of proof is (almost) always on the prosecutor. Earlier I spoke loosely of narratives offered by both the prosecution and the defense. The narratives offered by the defense, however, are generally parasitic on the narratives offered by the prosecution; they are often meant to weaken the prosecutor's case, without being full fledged alternative narratives. For this reason, I will mostly concentrate on the prosecutor's incriminating narrative. But what is a narrative?

7.2.2 Narrative construction

A crime narrative is progressively constructed, before and during trial, by the prosecutor, and in parallel, by the fact-finders. This construction is constrained by a number of factors: (i) the charges brought against the defendant; (ii) the items of evidence available prior to and during trial; (iii) common knowledge of what typically happens; and (iv) the envisioned or actual challenges coming from the defense. Given these constraints, we can view the construction of a narrative as a process driven by a dynamics consisting of 'giving rise to questions' and 'answering questions.' This process takes place in the mind of the prosecutor *before the trial*, but also in front of the fact-finders *during the trial*.

The process of narrative construction must begin with the following *composite principal question*: 'who did it? why? how? when? where?' The formulation of this composite question will heavily depend on (i) the charges brought against the defendant. After all, a murder case is different from a domestic violence case, or a pedophilia case, etc. But at the very minimum, the proposed narrative is expected to answer the principal question to some degree of detail. Now, as a narrative emerges in response to the principal question, other questions will emerge, and the latter, in turn, will have to be answered by a more refined narrative, and so on. The final narrative will be fully complete when all questions which could naturally or reasonably arise have been answered.

A narrative is not expected to answer all possible questions about what happened. Some questions will be reasonable and others far-fetched. This is mostly a function of the evidence available, item (ii) above, and of our common knowledge of what typically happens, item (iii) above. In the earlier example involving poison, asking whether an alien creature administered the poison to kill the victim would be far-fetched. But is it far-fetched to ask whether a friend of the defendant used the poison to kill the victim? The distinction between reasonable and far-fetched questions is not easy to draw. To gain some clarity, I suggest that we look more closely at the structure of narratives.

Following Pennington and Hastie (1991), let's define a narrative as a spatiotemporal configuration of events that are connected by relations of *physical causality* and *intentional causality*. A narrative is typically sub-divided into *episodes*, such as the one represented in the diagram on page 197. In each episode, we can identify a few constant and interconnected components: an initiating event; a resulting psychological response; goals, decisions and actions; and an outcome resulting from the actions being undertaken. Various episodes will be the building-blocks of a narrative. A narrative can be thought of as a hierarchy of embedded episodes, in which each episode fits in a coherent narrative structure, e.g. one episode triggers or enables the next one. Against this background, we can speak of a *bare narrative structure*—i.e. the minimal answer to the composite principal question—which is to be filled in with details according to ordinary patterns of physical and intentional causation.

Questions will naturally arise whenever a narrative is missing some of its parts. This can happen in at least two ways. For one thing, each episode in a narrative should have all of its parts, so questions will naturally arise whenever a psychological response is not preceded by any well-defined initiating event; whenever there is a mismatch between goals and action, or between action and its outcomes; whenever the purported relation of physical or mental causality is not specified; etc. These questions, I submit, will all be natural. For another thing, each narrative should connect, in the appropriate temporal and causal order, the different episodes that make up the narrative. If there are long and unexplained temporal gaps; if two episodes are not located in time; if the relation between two episodes is not specified; etc. These are all cases that would raise natural questions about the narrative.[7]

[7] These considerations are quite sketchy, and more work needs to be done. To think deeper about these issues, a starting point is the work by Hart and Honore (1985) who have considered the role of causation in the law, especially in civil cases. More recently, researchers within the AI & Law community have tried to come up with story-based ontologies for event reconstruction; see e.g. (Hoestra and Breuker, 2007). Note that the notion of 'ontology' referred to there is that of a computer scientist, not that of a philosopher;

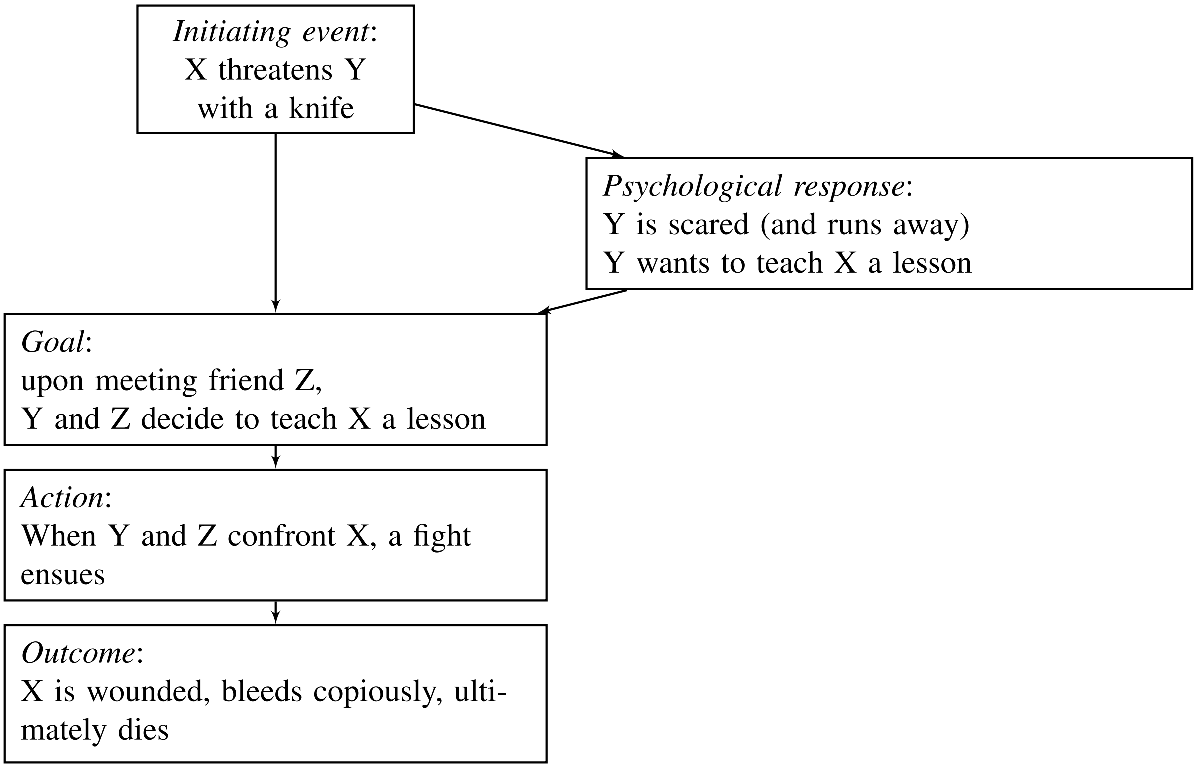

Initiating event:
X threatens Y
with a knife
Psychological response:
Y is scared (and runs away)
Y wants to teach X a lesson
Goal:
upon meeting friend Z,
Y and Z decide to teach X a lesson
Action:
When Y and Z confront X, a fight ensues
Outcome:
X is wounded, bleeds copiously, ultimately dies

Figure 7.2: Each episode typically consists of an initiating event, a psychological response, a goal, an action, and a resulting outcome.

Another factor that is helpful for drawing the distinction between far-fetched and reasonable questions is the actual or envisioned challenges coming from the defense, item (iv) above. To illustrate, suppose the defense endorses the hypothesis that a friend of the defendant administered the poison to the victim and offers suitable supporting evidence. The hypothesis in question, then, becomes salient, and consequently, the prosecutor has to take it seriously. On the other hand, if the defense disregards this alternative hypothesis, the prosecutor can happily set it aside. This is—to a certain extent—a merely tactical explanation of when a question is plausible or far-fetched, but I think that the dialectical confrontation at trial between the prosecution and the defense is not reducible to mere tactics. I shall say more on this topic in section 7.5.

7.2.3 Evidence-related narrative

In discussing the construction of a narrative, I spoke of questions regarding "what happened." But besides offerring a reconstruction of what happened, a narrative should also offer a reconstruction of how and why the proffered items of evidence bear on the case. In constructing a narrative, then, it is useful to distinguish between event-related questions and evidence-related questions. The principal question—who did it? why? how? etc.—is an example of an event-related question. I will now give some examples of evidence-related questions.

Suppose that certain testimony supports the conclusion that the defendant was running at a time and in a place close to when and where the crime was committed. This item of evidence seems certainly probative and material, and a narrative of the crime should contain details about when and where the defendant was running. In addition, however, the narrative should contain details about the evidence, e.g. whether the eyewitness was too far away or close enough to identify the defendant; why he was where he claims he was; etc. A narrative that failed to answer such evidence-related questions would certainly be incomplete, not because it failed to recount the events surrounding the crime, but because it failed to tell a proper story about the evidence itself. Consis-

see e.g. (Sowa, 2000).

tent with this observation, the F.R.E. explicitly require that a witness can testify only if 'evidence is introduced sufficient to support a finding that the witness has personal knowledge of the matter' (rule 102).[8]

Fingerprint evidence is no different. When fingerprints are found at the crime scene, the investigators must preliminarily assess whether they originated from the crime or not. Fingerprints do not speak for themselves; an argument is needed to show that they bear some relation to the crime. To this end, the investigators, or the prosecutor who wishes to introduce fingerprint evidence at trial, should reconstruct a spatiotemporally defined "sequence" which originates with the event of the crime and terminates with the fingerprint traces found by the police. Such sequence will contain spatiotemporal specifications tracking the history of the evidence, from the time of its origin up to its presentation at trial.

With these examples in mind, we can say that an evidence-related narrative should take the form of a "connecting sequence" which begins with the event of the crime; it describes how certain traces were formed as a result of the crime; and it ends with an account of how the traces were recovered by the investigators and were finally presented in court.[9] I am using the term 'traces' in a rather loose way, namely to refer to traces that are *physical* (e.g. fingerprints, blood stains, glass, hair, semen, etc.), *cognitive* (e.g. eyewitness' memories), and *digital* (e.g. recording from cameras, photographs, digital devices). The connecting sequence should be understood in causal terms; it must rely on the *causal expertise* of investigators, forensic experts, or whoever wishes to introduce an item of evidence.[10]

[8] This requirement is what some scholars have called *foundation*; see Schwartz (2011). In fact, if before or during trial it becomes apparent that nothing links the testimony to the crime, eyewitness evidence will be discarded, or at best, its probative value significantly curtailed. Recall the example from section 7.1: the defense lawyer asked the witness whether he played any team sport and it turned out that the witness identified one the people he played with as the perpetrator.

[9] In developing a general account of causality, Russell (1948) adopts a similar notion, that of a *causal line*. In a similar vein, Salmon (1984) develops an account of causality in terms of *continuous spatiotemporal processes*.

[10] Causal expertise consists in knowledge about causal relationships; see (Woodward, 2003) for a *manipulability* conception of causal relations; similarly, see (Lewis, 1973) for a *counterfactual* conception of causal relations.

Recall the fingerprints example. When the investigators or the prosecutor's forensic experts reconstruct the appropriate connecting sequence, they can do so on the basis of their existing causal expertise about how fingerprints persist through time, how they leave traces on different surfaces and materials, etc.[11] On this interpretation, the existence of a connecting sequence would be relative to the investigators' reconstruction and it would rest on their current and revisable causal expertise.[12]

***

Having explained what a crime narrative is supposed to be, I am finally ready to state the first requirement for the satisfaction of the criminal standard of proof:

> NARRATIVITY. The prosecutor should put forward a complete (or quasi-complete) incriminating narrative, i.e. a narrative that, at least, answers the principal question (who did it? why? how? etc.) and that, in addition, answers all (or most of) the event- and evidence-related questions which naturally arise.

Some might worry that this is too heavy a burden on the prosecutor, who cannot be expected to specify everything. I think this is a fair point and the choice ultimately rests with the prosecutor. Still, if he fails to address some of the questions that naturally arise from his proposed narrative, a good defense lawyer will seek explanations from the prosecutor. It is in the prosecutor's interest not to leave open too many "gaps" in the narrative, though some will be inevitable. More on this in section 7.5. In the next

[11] Some may wonder why the sequence must be causal and why it cannot be a series of spatiotemporal specifications about where and when the evidence originated and about how it persisted until the investigators found it. My sense is that spatiotemporal specifications would be insufficient, and it is inevitable to take recourse to causal patterns.

[12] The sequence, then, would be causal in an epistemic/subjective rather than ontological/objective sense, and as a consequence, it would be *subject to revisions*. For the investigators and the prosecutor might initially insert a piece of evidence in an appropriate sequence, but further scrutiny might prove the connection to be spurious. A sequence can be shown to be spurious in many ways. Consider the case of fingerprints: fingerprints might not match; they might have been planted; they might have been altered over time; they might have been left innocently; laboratory analyses might have been mistaken; etc. Thomson (1986), in contrast, seems to have a more ontological reading of the causal sequence.

section, I will supplement the narrative completeness requirement with a further one, i.e. that narratives be well supported by the evidence.

7.3 HIGH PROBABILITY ON THE EVIDENCE

Cognitive scientists have always been interested in the role that invariances, causal relations, and unifying patterns play in our understanding of the world (Sloman, 2005). Relatedly, there is now a body of legal and psychological literature suggesting that, as they search for these regularities and invariances, fact-finders construct their own crime narratives from the evidence (Pennington and Hastie, 1991; Thagard, 2000; Simon, 2004; Griffin, 2013). Even if the prosecutor, paradoxically, decided to give to the fact-finders the bare evidence alone, without providing any unifying narrative of the crime, the fact-finders will most likely make sense of the evidence by constructing their own narrative.

What I've been suggesting so far, then, is in line with current psychological findings insofar as it rests so centrally on the notion of a narrative. But the criminal justice system would not be in good shape if the party who prevails is simply the one who can tell the better story. It might very well be true that 'it would hardly shock lawyers ... to discover that the winner in some trials is the more sophisticated and compelling story teller' (Weisberg, 1996, p. 64). But this discovery would shock those who think that trials should be about finding facts on the basis of the evidence, not about about making things up. I am among those people. Criminal trials—I think—should not turn into competitive story-telling. To ensure this, we can view the BARD standard as a requirement that the prosecutor's narrative be strongly supported by the evidence. We can state this requirement in probabilistic terms, as follows:

> HIGH PROBABILITY. The prosecutor should put forward an incriminating narrative that is highly probable on the evidence.

A question arises: what does it mean that a narrative is highly probable on the evidence? The legal probabilists won't have too much trouble with the idea of a narrative being more or less probable on the evidence, but others might wonder whether this ideas can be made sense at all.

7.3.1 Four relations of evidential support

To begin with, I want to examine, in a probability-neutral way, how a body of evidence *epistemically supports* a crime narrative. I do so by distinguishing four cases:

> SIMPLE SUPPORT. A piece of evidence e supports proposition p as part of narrative N.
>
> CORROBORATION. Pieces of evidence e_1 and e_2 are independent and both support the same (or roughly the same) proposition p as part of narrative N.
>
> COMPLEMENTARITY. Pieces of evidence e_1 and e_2 support propositions p_1 and p_2, respectively; both propositions p_1 and p_2 are part of narrative N.
>
> INTERLOCKING. A piece of evidence e_1 supports proposition p_1 on the assumption that p_2 is true, and evidence e_2 supports proposition p_2 on the assumption that p_1 is true; both propositions p_1 and p_2 are part of narrative N.
>
> FILLING-IN. Pieces of evidence e_1 and e_2 are independent; they support propositions p_1 and p_2, respectively; the conjunction of p_1 and p_2 supports p, although e_1 or e_2 independently do not; the three propositions, p_1, p_2, and p, are all part of narrative N.

We've seen an example of *simple support* earlier in section 7.1. Recall the example of DNA expert testimony. The expert's assertion that the defendant's genetic profile matched with the semen on the victim's body was evidence for—i.e. it evidentially supported—the proposition that the defendant had intercourse with the victim. The relation of evidential support, as is typically the case, relied on generalizations and background knowledge of various kinds.

And now, an example of *corroboration*. Suppose that a witness says that the defendant was in Union Square in San Francisco at 5:30 PM, and suppose that another witness says roughly the same. This is a case of corroboration, for the two witnesses corroborate each other by saying the same thing. But what is the 'same' thing? Some degree of flexibility is needed here. A witness might say that the defendant was in Union square at 5:30 PM and another might say that the defendant was in Union Square at 5:40 PM. It is not clear whether this is an instance of corroboration. It depends on how crucial it is for the overall case to determine the exact time. If ten minutes make a difference and can exculpate the defendant, then the two witnesses are anything but corroborating each other. This suggests that the notion of corroboration itself cannot be analyzed by simply looking at two pieces of evidence in isolation. Corroboration has to be analyzed by considering the proposed narrative as a whole and its role in the prosecutor's case.

Next, an example of *complementarity*. Suppose witness 1 says (A_1): the defendant was at the station at 6:00 PM in San Francisco. Suppose that witness 2 says (A_2): the defendant was in San Jose at 8:00 PM. The two witnesses complement each other insofar as A_1 and A_2 are parts of the same narrative. The more connected—causally or temporally—the two parts of the narrative, the stronger the complementarity of the two pieces of evidence. Note that we have a case of complementarity here only assuming that the prosecution or the defense are offerring a narrative describing the defendant's whereabouts between 6:00 PM and 8:00 PM in the Bay Area. If the proposed narrative were only concerned with what happened in the surroundings of the San Francisco train station, we would have no complementarity at all. The notion of complementarity, then, is once again narrative-dependent.[13]

Interlocking is a stronger variant of complementarity. Consider again the proposition (A_1) that the defendant was at the station at 6:00 PM in San Francisco, and the proposition (A_2) that the defendant was in San Jose at 8:00 PM. Now, proposition A_1 alone does not support the conclusion (C_1) that the defendant was about to board the train going to San Jose; after all, being around a train station is no strong evidence that one is about to take the train. However, on the supposition that (A_2) is true, then (A_1) becomes better evidence for C_1. Similarly, proposition A_2 alone is weak evidence for

[13] As another illustration of the same point, suppose a witness testifies that the defendant is a sales manager in a department store and another witness testifies that the defendant was in Union Square in San Francisco at 5:30 PM. Do the two items of evidence constitute a case of complementarity? It depends on the overall narrative; it depends on whether they support certain propositions which are both part of a narrative being proposed.

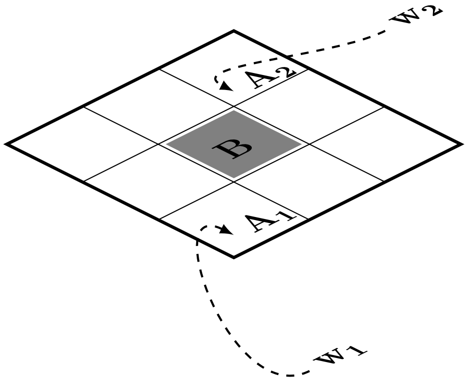


Figure 7.3: A pictorial illustration of *filling-in*. The proposition B fills in the gap between A_1 and A_2, which are respectively supported by item of evidence w_1 and w_2. Instead, B is not supported by any item of evidence in particular.

the conclusion (C_2) that the defendant took the train from San Francisco, but on the assumption that A_2 is true, A_1 becomes better evidence for C_2.

Finally, an example of *filling-in*. Consider again propositions A_1 and A_2. Their conjunction supports the proposition (B) that the defendant traveled by train between 6:00 PM and 8:00 PM from San Francisco to San Jose. Importantly, by adding B to A_1 and A_2, the overall completeness of the narrative increases because the question where the defendant was between 6:00 PM and 8:00 PM finds its natural answer. Note that proposition B is supported by witness 1 and witness 2 only indirectly; it is supported by the two testimonies *through* the mediation of the narrative.[14]

7.3.2 Cohesiveness

An analogy from Susan Haack (1995, 2011) is useful here. Haack likens our epistemic inquiries to solving crossword puzzles. In our case, the parallel is between solving a

[14] Here is another example of filling-in, with three pieces of evidence. Suppose that, first, a cut on victim's body suggests that (p_1) X killed victim with a knife; second, semen on the victim's body suggest that (p_2) Y had intercourse with victim; and finally, a DNA analysis of the semen suggests that (p_3) the defendant is linked to the crime scene. The three pieces of evidence, through the mediation of a narrative, supports the proposition that (p_4) X and Y are the same person and that this person is the defendant.

crossword puzzle and constructing a crime narrative. When one attempts to solve a crossword puzzle, it is as though one is putting together the different pieces of a crime narrative. There are different clues, each supporting certain entries; similarly, there are different pieces of evidence, each supporting certain propositions as parts of a narrative. The clues together won't typically be enough to fill out the entire crossword puzzle, and likewise, the available items of evidence will typically not be enough to support each part of the narrative. (And, incidentally, this is often more so with crime narratives than with crossword puzzles). Some filling-in—or some educated guessing—will then be necessary.

Filling-in and interlocking are ways to invent and make up one's narrative when no direct supporting evidence is available for certain parts of it. Yet, they should not be done arbitrarily. They should be the result of a *cohesiveness-based reasoning*, a type of reasoning that is driven by the need to eliminate gaps in a narrative, on the one hand, while keeping up with what is most probable given world knowledge and given the others parts of a narrative. If there is e.g. a temporal gap in a narrative, considerations based on cohesiveness can help us fill it in by adding whatever is most probable given common sense and given the other parts of the narrative.

7.3.3 Evidential support and probability

The foregoing discussion illustrates that narratives are not mere fictions so long as suitable evidential constraints are in place: simple support, corroboration, complementarity, and filling-in. My discussion of the four relations of evidential support has refrained from using probability, but the legal probabilists must be curious whether the four relations can be formulated in a probabilistic way. It should be possible, although it is not a trivial task. Let me sketch some remarks on this score.

For sure, both 'simple support' and 'corroboration' can have a probabilistic rendering. A piece of evidence would support a proposition as part of a crime narrative if the evidence increases the probability of the proposition (Fitelson, 2006; Lempert, 1977). Corroboration can be represented probabilistically as a phenomenon of "confidence boost" so that the probability of a proposition, given two corroborating items of evidence, would be greater than the probability of the same proposition given one item of evidence only (Bovens and Hartmann, 2003). This idea is expressed quite simply by Cesare Beccaria:

> When the pieces of evidence are independent of each other, that is, when the evidence can be tested other than by each other, then, the more evidence is adduced, the more credible [probable] is the matter in question. (Beccaria, 1764, Cp. 14).

It is more difficult to provide a probabilistic formulation of the other three relations: complementarity, interlocking and filling-in. And to my knowledge, there is no probabilistic treatment of such relations in the literature. I will not provide one here, but in what follows, I shall assume that a probabilistic treatment can indeed be provided. Once this treatment is given (and granted some dose of idealization), talking of a narrative's probability on the evidence should appear less unpalatable.

7.4 EVIDENTIAL COMPLETENESS

So far I have taken for granted that a determination of guilt must be based on the evidence. This is not incorrect, but it is complete. What fact-finders should consider for the purpose of their determinations is the presence *as well as* the absence of evidence. The Supreme Court of Canada in R. v. Lifchus (1997) is adamant that a reasonable doubt is 'logically connected to the evidence or absence of evidence.' A probabilist like David Kaye (1986b) recognizes the epistemological relevance of gaps in the evidence, and he writes that 'a jury will expect to hear certain items of evidence in certain cases, and it may regard the failure to produce such evidence with devastating skepticism' (p. 663). This is a fact about the psychology of jurors and also an epistemically important fact. In the legal fact-finding process, what matters is the available evidence *as well as* how complete or incomplete such evidence is. The more complete the incriminating evidence, the better; the more gaps in the evidence, the weaker the incriminating case. But when is the evidence complete or incomplete?

7.4.1 Expectations

In a general sense, the evidence for a conclusion or a narrative is always incomplete. The completeness of the evidence can only be so in a limited, restricted, and relative way. But relative to what? To a first approximation, the evidence is complete or incomplete relative to what the fact-finders can reasonably expect. What we need, then, is an account of when the fact-finders can reasonably expect to see more evidence and when they should not.

In ordinary life, what evidence to expect in support of a certain conclusion is often a straightforward matter. Suppose I am holding a lottery ticket and I wonder whether my ticket is the winner or not. In this situation, I can reasonably expect that there will be a lottery drawing and that the result will be published in a newspaper. No mystery about that. According to the *normal, expected path of evidence acquisition*, there will be a lottery drawing and the result will be in the news. If I fail to be informed about the lottery result, my evidence for whether or not my ticket is the winner will be incomplete. (For more on this, see chapter 6.) Something similar must be going on with the fact-finders in a criminal trial. For instance, in a drunk driving case, the fact-finders will expect testimony about the defendant's alcohol level and the use of a breathalyzer; if no mention of that is made, the fact-finders will be perplexed.

Our expectations depend on two factors. One factor is our societal and common knowledge regarding what types of evidence are available. Thirty years ago no one would have expected to see DNA evidence in a rape case, but today fact-finders do, and they will want to hear an explanation if DNA evidence is not presented. The second factor that affects our expectations is the probandum in support of which evidence is being offerred. Recall from section 7.1 that evidence is a relational notion between a probans and a probandum; recall also that various probanda typically fit together in a larger narrative. I am going to suggest that the proposed narrative affects our expectations a great deal.

7.4.2 Narratives and gaps

A narrative of the crime is *supported* by the evidence, but it also *explains* or *accommodates* the evidence (Pardo and Allen, 2008). The relationship between evidence and narrative goes both ways: from the evidence to the narrative (evidential support) and from the narrative to the evidence (explanation, accommodation). If the perpetrator used a knife to kill the victim (narrative), this explains why the victim had a cut on her throat (evidence), and vice versa, the cut on the victim's throat (evidence) supports the conclusion that the perpetrator used a knife to kill (narrative).

As a consequence of its explanatory power, when a narrative is believed or at least entertained, it creates *expectations* regarding which items of evidence are to be presented or not during the trial. Such expectations stem from reasoning along the following lines: if the prosecutor's narrative says that a certain event occurred, the latter must have left certain traces at the crime scene; careful police investigation should have identified those traces and presented them as evidence during the trial proceedings. For instance, if a narrative says that the perpetrator killed the victim after sexual intercourse, the fact-finders will expect that appropriate crime traces be identified at the crime scene and thus be discussed by a forensic expert at trial; if a narrative says that the defendant kept a detailed diary, the fact-finders will expect a diary to be found which contains a recounting of what happened the day of the crime; if a narrative says that the crime happened in a crowed place, the fact-finders will expect some witnesses testifying; etc. These examples suggest the following account of gaps in the evidence:

> TYPE ONE GAPS (OR GAPS OF MISSING PREDICTION). There is a gap in the evidence whenever the prosecutor's narrative creates the expectation that evidence E be presented at trial, but E is not presented.

On the present account, the fact-finder's expectations about the evidence to be presented at trial are anchored to the prosecutor's incriminating narrative. The body of evidence, then, will count as complete or incomplete only relatively to a proposed narrative, not absolutely. On an absolute scale, after all, the total evidence presented is always going to be incomplete.

(Incidentally, there is an added layer of complexity here. It is important to realize that the fact-finders are always given a limited body of evidence—i.e. the total evidence as presented at trial—and they are asked to reach a decision on the basis of that evidence, and that evidence only. It is as though the fact-finders were told: *reach a decision on this evidence as if it were the only evidence one could conceivably have, i.e. turn off your expectations of further evidence*. Many items of evidence might have been stricken from the record because they were inadmissible. Further, fact-finders, and jurors in particular, are not allowed to do their own research or seek new evidence. Think of the peculiar situation in which jurors are placed: they are locked in a room, isolated from the rest of the world, and asked to reach a verdict only on the basis of common sense and of the evidence presented at trial. The legal decision-making process is an extremely regimented process. Inevitably, if certain items of evidence are declared inadmissible, the fact-finders will find themselves in a difficult predicament. Suppose an impressive statistic about a genetic profile's frequency is stricken from the record because of its prejudicial effects, or suppose some photographs are stricken from the record because they are too gruesome. The fact-finders might be left wondering: Is the evidence missing because it was, in fact, missing or because it was declared legally inadmissible? If they infer that the evidence was, in fact, missing, this will be to the detriment of the prosecutor's case, at least more so than had they concluded that the evidence existed but was inadmissible.)

Besides type one gaps, there is another possible type of gap in the evidence. Suppose that the prosecutor's narrative is that the defendant was on a train from San Francisco to San Jose between 4:00 PM and 6:00 PM. Naturally, the fact-finders would expect the prosecutor to offer certain items of evidence here, e.g. that the defendant was seen in San Francisco before 4:00 PM and in San Jose after 6:00 PM; that the defendant was seen in the vicinity of the San Francisco and San Jose stations; that the defendant was seen on the train between 4:00 PM and 6:00 PM; etc. There would be gaps in the evidence if such or similar items of evidence were missing. The reason is that certain parts of the narrative would be unsupported by direct evidence. More generally:

TYPE TWO GAPS (OR GAPS OF MISSING SUPPORT). Whenever there is

> no direct evidence supporting some portions of the prosecutor's narrative, there will be gaps in the evidence.

Building on the foregoing discussion, I can now offer a further requirement on the BARD standard:

> EVIDENTIAL COMPLETENESS. The prosecutor should make sure that, *relative to the proposed narrative*, the evidence presented at trial is complete, i.e. it is not affected by type one or type two gaps. If there are type one gaps in the evidence, the prosecutor should give a satisfactory explanation for them. If there are type two gaps in the evidence, the prosecutor can "fill in" the narrative (see the previous section for an account of filling-in).

Type one gaps are particularly hard to explain away. If the narrative, for example, creates the expectations that certain crime traces should have been found at the crime scene and presented as evidence at trial, whenever the expectation is not fulfilled, the prosecutor had better adjust his narrative accordingly to account for the failure to present those traces. Type two gaps are easier to explain. Given that stories are complex structures of events, the fact-finders cannot expect that the prosecutor be able to offer evidence for each specific part of the narrative. It is enough for the prosecutor to offer supporting evidence for the main parts of the narrative, while the other parts are the result of "filling in" the gaps.

7.5 RESILIENCY

The account of BARD I formulated so far consists of three components: narrativity; high probability of the narrative given the evidence; completeness of the evidence. From these three requirements alone, it looks as though the satisfaction of the criminal standard of proof should rest exclusively on the prosecutor. Although the burden of proof is on the prosecutor, criminal trials are adversarial and the defense should have an opportunity to challenge the prosecutor's case. This observation suggests a further requirement:

> RESILIENCY. The prosecutor's narrative should resist all challenges coming from the defense lawyer who had an effective opportunity to scrutinize the prosecutor's narrative and who took full advantage of such an opportunity.

7.5.1 Less than full resiliency.

In probabilistic terms, the resiliency condition amounts to the requirement that the prosecutor's narrative be highly probable even in light of challenges and contradictory evidence. Let's now turn to legal resiliency. First of all, probability-minded epistemologists well admit of a difference between 'high probability' and 'resiliently high probability' (Skyrms, 1977, 1980; Leitgeb, 2010). Suppose the probability that it rained, given that the road is now wet, equals 0.9, because e.g. experience tells us that, in 1 out of 10 cases, a wet road follows the rain. Though 0.9 is a high probability, you might wonder whether it can change significantly in light of future evidence or alternative explanations. For instance, consider the hypothesis that the sprinklers have been watering the grass on the sidewalks. This hypothesis, if true, would bring down the value of the initial probability assignment. The probability than it rained, given that road is now wet and—*in addition*—that the sprinklers have been watering the grass on the sidewalks, must be significantly lower than the initial 0.9 value. So, this value was high but not very resilient. Note that the resiliency of a probability value assignment is always relative to a proposition (or a set of propositions) describing an item of evidence or an hypothesis. In our example, the value 0.9 was not very resilient relative to the proposition 'the sprinklers have been watering the grass on the sidewalks.'

Let's state all this more generally. Suppose the probability of a proposition A, given a body of evidence E, equals some value r between 0 and 1; we write $P(A|E) = r$. The resiliency of a statement that assigns a probability value to a proposition (a probability statement, for short) can be formulated as the complement of the statement's variability, so that the more variability, the less resiliency. Now, the variability of $P(A|E) = r$, relative to a proposition π_i, is the absolute difference between r and the probability of A given E *and* π_i, or in short, $|r - P(A|E \wedge \pi_i)|$. Variability can be defined more generally over a set of propositions $\Pi = \{\pi_1, \pi_2, \pi_3 \dots\}$ by taking the widest variability, namely $max_i\{|r - P(A|E \wedge \pi_i)|\}$, where π_i is any proposition in Π such that π_i and E are consistent. (E and π_i must be consistent because the conditional probability of a proposition given an inconsistent proposition is typically undefined.) Following Skyrms (1980), a definition of resiliency can now be stated:

> The SKYRMS RESILIENCY of a conditional probability statement $P(A|E) = r$, relative to a set of propositions Π, is given by 1 *minus* its variability, i.e. $1 - max_i\{|r - P(A|E \wedge \pi_i)|\}$. (If the variability of a conditional probability statement is zero, then its Skyrms resiliency will be complete, as expected.)

The Skyrms resiliency of a probability statement comes in degrees between 0 and 1, and it is always relative to a set of propositions Π. Full and absolute resiliency—i.e. resiliency with degree one relative to all the propositions in a language—is very hard to come by. One question that naturally arises here is, how are we to select the set Π? This question, as we shall see, is particularly pressing if we want to apply Skyrms resiliency to the trial context.

The notion of resiliency thus described can be applied—*mutatis mutandis*—to the trial context. We can offer an account of the resiliency of probabilistic claims of the form $P(N|E) = r$, where N stands for the prosecutor's narrative and E denotes the available evidence (at some point during the trial). In criminal trials, the set of propositions relative to which the probability of guilt can vary—that is, the set of propositions relative to which resiliency is measured—will consist of the challenges, objections, and counter-evidence which the defense can present during its scrutiny of the prosecutor's case. Let Σ denote the set of propositions describing challenges, objections, and counter-evidence. The propositions in Σ will be of two types: items of exculpatory evidence and exculpatory hypotheses about what happened. Now, Skyrms resiliency—see previous footnote—does not distinguish between upward or downward variability, since it uses the absolute value of the difference. In the context of criminal trials, however, we are mostly interested in downward variations of the guilt probability on the evidence. The problematic case arises when the guilt probability goes *down*, maybe even below the threshold guilt probability. Legal resiliency, then, is a subcase of Skyrms resiliency, as follows:

> The LEGAL RESILIENCY of a conditional probability statement $P(G|E) = r$, relative to a set of propositions Σ, is given by 1 *minus* its downward variability, i.e. $1 - max_i\{|r - P(A|E \wedge \pi_i)|\}$ restricted to only the π_i's such that $r \geq P(A|E \wedge \pi_i)$.

If legal resiliency is part of the criminal standard, what would be the appropriate degree of resiliency that is required to convict? A suggestion here is that the variations in the probability—relative to the propositions in Σ—should never go below a fixed probabilistic threshold in accordance with the requirement of evidential support. Legal resiliency would thus require that, even taking into consideration the exculpatory evidence and hypotheses in Σ, the probability of the prosecutor's never should never go below a fixed threshold.

Another question arises. Which propositions should be admitted, so to say, in Σ and which propositions should be left out? Clearly, if any proposition were admitted, even far-fetched speculations, no probability assignment would ever be legally resilient. We should make sure that legal resiliency does not become too demanding a condition. In searching for some criteria to identify the Σ-admissible propositions, we can do no more than appeal to reasonableness. Which exculpatory hypotheses are reasonable and which are not? Which items of exculpatory evidence could reasonably emerge in the future? Answering these questions, I think, is not a matter that can be decided beforehand once and for all. The reasonableness of an alternative hypothesis primarily depends on what prosecution and defense disagree or agree upon. It is through the trial dynamic of questions and answers that an alternative hypothesis, even a far-fetched one, can become reasonable. While crimes follow certain patterns, they are also filled with unexpected turns of events. In principle, any alternative hypothesis could be a reasonable one, so long as the defense makes a *prima facie* case for it. And if that alternative hypothesis can lower the guilt probability below the threshold value, this would be a good enough source for a reasonable doubt.

The criminal standard hints at the idea that the prosecutor's case should survive the challenges levelled against it. Establishing a case beyond any reasonable doubt, after all, means establishing it *beyond* any reasonable challenge that is levelled against it. The resiliency condition I've inserted in the formulation of BARD captures this idea. Resiliency should be relativized to trial proceedings and to the adversarial process—to what we might call, broadly speaking, cross-examination. This is a process by which the defense tests whether the prosecutor's case can withstand challenges and objections. If the defense's cross-examination does not shake the prosecutor's case, resiliency would be satisfied. In probabilistic terms, resiliency would be satisfied whenever no reasonable defense objection—be it actual or potential—could lower the initial guilt probability.

7.5.2 Effective opportunities

As I understand it, the resiliency condition requires that the defense have an *opportunity* to level challenges against the prosecutor's narrative and that it take full advantage of such an opportunity. If the defense were inert, or unable to issue challenges, the resiliency condition would not be satisfied, and the criminal standard would not be satisfied either. In addition, it is important that the defense is given, not just the mere opportunity, or a purely formalistic and idle opportunity, but an *effective* opportunity to scrutinize the prosecutor's case. It is only if the prosecutor's case can withstand a vigorous, extensive, powerful cross-examination that the fact-finders can be in a position to expect that if new evidence were to come up, it would not shake the prosecutor's case in any significant way (Stein, 2005).

What is an effective opportunity to scrutinize the prosecutor's case? This is a particularly difficult question. In many countries there are specific constitutional guarantees that ensure that defendants have enough monetary, intellectual, and evidentiary resources to exercise their right to a defense. In the United States, these guarantees amount to the right to counsel; the right to effective assistance of counsel; and the compulsory process. The *right to counsel* for indigent defendants is a way to make sure that even defendants with little monetary resources should be able to afford a counsel.[15] In addition, the *right to effective assistance of counsel* makes sure that defendants, not only have a counsel, but also have an effective counsel, one who complies with the intellectual and ethical standards of the profession.[16] Finally, the *compulsory process* gives the power to the defense to subpoena witnesses to testify in court. The rationale behind

[15] E.g. the right to counsel is protected by the 6th Amendement to the U.S. constitution; Gideon v. Wainwright, 372 U.S. 335 (1963), mandates that a counsel should be appointed for all indigent defendants in all felony cases.

[16] See McMann v. Richardson, 397 U.S. 759, 771 n.14 (1970); and more recently, Strickland v. Washington, 466 U.S. 668 (1984).

the compulsory process is that, if there is potentially exculpating evidence, defendants should be a position to make use of it in court.[17]

Ensuring that defendants have adequate monetary, intellectual, and evidentiary resources contributes to give them an effective opportunity to scrutinize the prosecutor's case. But is that enough? A defendant's resources should meet a minimum threshold, but they should also be on an *equal footing* with the prosecutor's resources. Whenever the prosecutor has disproportionately more resources, the defendant will not have an effective opportunity to level challenges, even though the defendant's resources might meet a minimum threshold. My suspicion is that in cases involving supposed expert witnesses testifying for the prosecutor on statistical issues, the defense can often fail to have the preparation, the expertise, or even the monetary resources that are required to scrutinize the statistical testimony properly.[18]

7.5.3 Challenges proper

The resiliency condition, as I've been articulating it, amounts to a resistance to challenges that are brought up by the defense when the latter has an effective opportunity to do so and takes full advantage of such an opportunity. We should now look more closely at the nature of these challenges or objections. The challenges can be directed at the supporting evidence or at the prosecutor's incriminating narrative. I am mainly interested in the latter. When the challenges are directed against the prosecutor's narrative, they can be of two types. They can target what the narrative says or they can target what the narrative does *not* say. I shall call the former *challenges proper*, and the latter *spoliation challenges*. Let's consider each in turn.

It turns out that a close correlation exists between a prosecutor's narrative and its susceptibility to challenges proper. Recall that narratives are structured sequences of events. Earlier I spoke of the completeness of narratives: a narrative is complete when-

[17] E.g. the *subpoena duces tecum* compels a witness to appear in court and disclose relevant documents. For a recent discussion of the compulsory process in U.S. law, see Hewett (2007).

[18] This is partly what happened in some of the cases mentioned in chapter 4. E.g. Sally Clark and Lucia de Berk were incriminated by means of flawed statistics, yet their defense teams failed to identify such flaws.

ever it has all of its parts and has no gaps; otherwise it is incomplete. For instance, if a narrative fails to identify the perpetrator, or if it fails to describe the initiating event, it will count as incomplete. More carefully, I should say that the completeness of a narrative is a matter of degrees: a narrative can lack a variable number of parts and can have more or less extensive gaps. We can never expect a prosecutor's narrative to be fully complete in all respects; something will inevitably be missing.

With this in mind, suppose a prosecutor's narrative in a homicide case is relatively incomplete: it does not describe the weapon used in the crime nor does it say how the victim was killed. To mount a challenge proper—i.e. a challenge directed against what the narrative says—the defense will have to target all the possibilities that the unspecified narrative leaves open. In this case, if the defense offers evidence that e.g. the killing did not happen through strangulation, the prosecutor could easily respond that he did not specify how the killing occurred, and that if the killing was not by strangulation, it was by some other, unspecified procedure. The challenges—to be at least *prima facie* successful—will have to address all possible ways the killing could have occurred, such as poisoning, strangulation, stabbing, and many others. All in all, a quite extensive body of arguments would be required to challenge—*prima facie*, let alone successfully—the narrative in question. In contrast, if the prosecutor's narrative specifically said that the perpetrator poisoned the victim, the defense would only need to formulate a challenge that was pertinent to poisoning; it would not need to consider other ways the perpetrator could have killed the victim.

I want to suggest the following correlation: a more complete narrative is more widely exposed to challenges because it says more, it commits itself to more propositions than a less complete narrative which leaves many issues and questions open. The correlation between the completeness of a narrative and its susceptibility to challenges echoes an idea of Karl Popper's. He believed that the hallmark of scientific hypotheses is that they are falsifiable, i.e. susceptible to be falsified by further observations (Popper, 1935, 2002). He also believed that a theory's susceptibility to falsification is a function of its informative content, of how much the theory tells us about the empirical world. In discussing the role of scientific hypotheses in science, Popper wrote:

> Science does not aim, primarily, at high probabilities. It aims at a high informative content, well backed by experience. But a hypothesis may be very probable simply because it tells us nothing, or very little. A high degree of probability is therefore not an indication of goodness. (Popper, 2002, p. 416, Appendix *IX)

This remark applies—*mutatis mutandis*—to incriminating narratives. The aim is not simply to offer highly probable narratives, but also narratives that can be adequately scrutinized and challenged. And I've suggested, in agreement with Popper, that there is a correlation between a narrative's degree of completeness and its susceptibility to challenges proper. The correlation matters a great deal for resiliency: between two narratives which both resisted challenges proper, the one which was exposed to more challenges should be regarded as overall more resilient than the one which was exposed to less challenges.

7.5.4 Spoliation challenges

Recall now the process of filling in a narrative. When a prosecutor's narrative is filled in with details and propositions that are not directly supported by the available items of evidence, this exposes the narrative to more potential challenges than if the gaps were left open in the narrative. Filling-in, then, is not only a weak form of evidential support, it is also a way to increase a narrative's exposition to challenges. This is a reason why an abuse of filling-in as a form of evidential support can lead to narratives that are weakly supported by the evidence and that can be easily exposed to challenges.

Leaving gaps in one's incriminating narrative, then, might seem a good strategy for the prosecutor. It is as though a prosecutor can reduce his exposure to potential challenges by leaving gaps in his narrative. This would be true if a crime narrative could only be challenged through what I call *challenges proper*, i.e. challenges to what the narrative says. But a narrative can also be challenged for what it does not say. I call this second type of challenges *spoliation challenges*. If a prosecutor does not localize a time and place for the crime, or if she refrains from telling how the crime occurred, whether a weapon was used, how many people were involved, etc., all these gaps or "silences" would be the target of the defense's cross-examination.

It is a question of prosecutorial strategy which gaps can be left open in the incriminating narrative, while being wary that the defense might raise spoliation challenges. Prosecutors here face a dilemma. If they leave manifest gaps in their incriminating narrative, the fact-finders might draw negative inferences or the defense might raise spoliation challenges that could indirectly weaken the prosecutor's case. On the other hand, if prosecutors fill in the gaps, they will expose their narratives to more challenges proper, thereby risking weakening their case. This complex dynamics between susceptibility to proper and spoliation challenges, on the one hand, and filling in and leaving open gaps, on the other, is unique to the dialectical and adversarial aspect of criminal trials. The resiliency condition in my formulation of BARD is a pointer toward this dynamics. It is a dynamics—I want to emphasize—which becomes intelligible only if we assume that what prosecutors are doing is to offer well-specified and cohesive narratives in which gaps can be either filled in or else properly explained.

An acceptable gamble? There is a dose of gambling in the prosecutor's strategy. The prosecutor can decide to leave gaps in the incriminating narrative, hoping that the defense will not raise spoliation challenges. Alternatively, he can fill in the gaps, hoping that the defense will not raise challenges proper. The prosecutor, loosely speaking, engages in some form of gambling here—but it is an acceptable one. He is not gambling with the defendant's life or liberty. He is gambling with how to formulate the incriminating narrative. The prosecutor's gamble is not the end of the story. The defense still has the opportunity to respond.

7.6 BARD UNRAVELED

My explication of the criminal standard is now complete. To recapitulate, here is my four-pronged account:

> When a prosecutor attempts to establish guilt beyond a reasonable doubt, four requirements apply:

> NARRATIVITY. The prosecutor should put forward a complete (or quasi-complete) incriminating narrative, i.e. a narrative that, at least, answers the principal questions (who did it? why? how? etc.) and all (or most of) the event- and evidence-related questions which naturally arise.
>
> HIGH PROBABILITY. The prosecutor should put forward an incriminating narrative that is highly probable on the evidence.
>
> EVIDENTIAL COMPLETENESS. The prosecutor should make sure that, *relative to the proposed narrative*, the evidence presented at trial is complete, i.e. it is not affected by type one or type two gaps. If there are type one gaps in the evidence, the prosecutor should give a satisfactory explanation for them. If there are type two gaps in the evidence, the prosecutor can "fill in" the narrative (see the previous section for an account of filling-in).
>
> RESILIENCY. The prosecutor's narrative should resist all challenges coming from the defense lawyer who had an effective opportunity to scrutinize the prosecutor's narrative and who took full advantage of such an opportunity.

7.7 PRISONERS

I now want to put my proposal to a test by using a well-known legal hypothetical. The hypothetical has been used as an argument that a probability-based account of the BARD standard gives rise to counterintuitive results.

> *Prisoners*. A video recording shows that, in a prison yard, ninety nine out of one hundred prisoners assaulted and killed the only guard on duty. In the recording, ninety nine prisoners appear to deliberately participate in the killing; only one refrained. The faces are unrecognizable, and no other evidence is available. All prisoners escaped except one, Eschaton, who is captured and tried. Given the 99:100 statistical rate of participation in the killing, one might argue that Eschaton's guilt is 0.99 probable and that Eschaton should be convicted because a high probability of guilt is enough to satisfy the criminal standard of proof. This argument, however, would be considered unacceptable by nearly everybody. [19]

[19] This is a modification of a scenario in (Nesson, 1979, p. 1192).

Most of us have the gut reaction that it would be unacceptable to convict Eschaton, even if his probability of guilt is high.[20] The interesting question is whether my proposed explication of BARD agrees with the intuition that convicting Eschaton would be unacceptable. Although we are dealing with a quite artificial scenario, it is worth asking what charges could possibly be brought against Eschaton. I shall work on the assumption that Eschaton has been charged with some kind of homicide.

To begin with, note that the prosecutor failed to offer a well-specified narrative of the crime. Recall section 7.2. Following Pennington and Hastie (1991), there I defined a crime narrative as a spatiotemporal configurations of events that are connected by relations of *physical causality* and *intentional causality*. To offer an adequate reconstruction of the concomitant occurrence of *mens rea* and *actus reus*, a crime narrative should contain a description of e.g. initiating events; psychological responses and goals resulting from those initiating events; actions, consequences and accompanying states resulting from the actions. In *Prisoners*, what is the initiating event? What psychological responses did it trigger? Who participated in the killing? What were the different participants doing? If Eschaton participated in the killing, what was *he* doing? What prompted him to go kill the guard? To all these questions, the prosecutor's narrative—whatever it may be—offers a vague or no answer at all. So, the first observation to make here is that the prosecutor's incriminating narrative is grossly incomplete. (To be sure, the video recording gives a fairly detailed reconstruction of what happened. From the video, in all likelihood, we can reconstruct what each single prisoner did. Yet—and this is where the narrative becomes grossly incomplete—the recording fails to offer a reconstruction of who did what, and most importantly, the recording fails to offer a reconstruction of what Eschaton, among the other prisoners, did.)

Some commentators observed that the statistical evidence against Eschaton is not individualized or specific enough. The idea here is that if the statistical evidence in *Prisoners* incriminates Eschaton, by parity of reasoning, it incriminates any other prisoner. In this sense, the statistics are not specific enough to discriminate a prisoner from

[20] Wells (1992) tested people's reactions in scenarios similar to *Prisoners*; most people feel uneasy in convicting, *even when* the governing standard of proof is lower than *beyond any reasonable doubt*.

the other; they are not specific enough to identify the prisoner who did not participate in the killing (Nesson, 1979; Pundik, 2011).[21] I prefer to focus on the deficiencies in the incriminating narrative, rather than on deficiencies in the incriminating evidence, although the two are strongly related.

So, the first reason why Eschaton's guilt has not been established BARD—I suggest—is because the incriminating narrative is grossly incomplete. The second, closely connected reason has to do with the resiliency condition. Given certain assumptions, Eschaton's probability of guilt is high, yet it might not be resiliently high. Imagine new evidence becoming available, such as DNA traces on the victim's body, fingerprints, or a witness. Then, it might very well turn out that Eschaton was innocent. Eschaton's probability of guilt is not particularly stable nor resilient; it is high just because the evidence is extremely limited. So, the resiliency condition fails.

Some will object that, in every criminal case, new and surprising evidence could always undermine the prosecutor's case, yet this possibility alone should not be a reason to think that guilt was not established beyond a reasonable doubt. I agree. Otherwise criminal convictions would become impossible. In fact, the resiliency requirement I've proposed for the criminal standard is not that of full resiliency, but that of resiliency limited to the challenges that the defense would raise under suitable conditions. My point is not that Eschaton's probability of guilt is not resilient as such. More carefully, I should say that since the prosecutor did not offer any well-specified narrative, it is hard to say whether the defense could have any effective opportunity to raise challenges against it. Without the defence being in a position to level challenges, it is difficult to say whether the prosecutor's case is resilient or not. In order to decide whether guilt has been proven beyond a reasonable doubt, we would need to see what challenges the defense can raise and how the prosecutor responds.

My conjecture is that the defense can raise no challenges proper, for the prosecutor's narrative is so grossly incomplete and its susceptibility to challenges proper is very lim-

[21] Goldman (1976), Lewis (1996), Hawthorne (2004), and Kvart (2006) also make a similar point while discussing scenarios that resemble *Prisoners* (i.e. lotteries in which all tickets lose except one, or villages in which all barns are real except one).

ited. (On this score, recall the correlation between a narrative's degree of completeness and its susceptibility to challenges proper.) Consequently, the defense will mostly raise spoliation challenges that attack the gaps in the prosecutor's narrative. I do not see how the prosecutor could adequately respond to the spoliation challenges except by gathering new evidence or by filling in his narrative in the appropriate way. But as soon as the prosecutor does so, we would have a much different case from *Prisoners*.

In short, I think there are two, closely interconnected reasons that justify our reluctance against convicting Eschaton. The first is the absence of a well-specified narrative. The second is that the prosecutor's narrative fails to be resilient.

7.8 MULTIPLE QUESTIONS

The account of the criminal standard I've proposed consists of four parts: narrativity; high probability (of a narrative, not of guilt simpliciter); evidential completeness; and (suitably restricted) resilience against challenges. These parts closely parallel four different questions that should be asked in criminal trials:

> (*factual*) Is the defendant guilty or not?
>
> (*evidential support*) Does the evidence supports guilt or innocence?
>
> (*evidential gaps*) Is the evidence presented at trial complete or not?
>
> (*dialectical*) Is a putative decision toward guilt robust (or resilient) enough to resist challenges?

Ideally rational fact-finders can form reasonable doubts as they entertain any of these four questions. A reasonable doubt can arise from possible gaps in the incriminating narrative; from whether or not the evidence adequately support the incriminating narrative; from gaps in the evidence; from how easily the incriminating narrative could be overturned by other evidence.

As Charles Nesson (1985) remarked, a probabilistic interpretation of the BARD standard focuses exclusively on the question of evidential support. On this interpretation, a reasonable doubt is solely associated to whether or not guilt is sufficiently probable on the evidence. Yet, Nesson thinks, convictions should not present themselves to be about the evidence, or to be about the relation of the evidence to the facts; they should present themselves to be about the facts. As he puts it:

> The grammar and logic of judicial proof . . . suggest that the very essence of the rule of law is the idea that legal sanctions should be imposed on the basis of rules applied to conclusions about what happened, not to conclusions about the evidence. (Nesson, 1985, p. 1391)

If this is correct, I think that the central role that narratives play in my account can vindicate the intuition that verdicts should be about facts and not about the relation of the evidence to the facts.

But criminal trials are not simply about finding the facts. The conclusions reached about the facts should be tenable and justified, on the one hand, and the parties involved should have reached such conclusions through their best possible efforts. What does that mean? We are used to thinking of the criminal standard as an *epistemic* device that measures the strength of the evidence, probabilistically or non-probabilistically (Shapiro, 1991). But the criminal standard has also a strictly moral flavour, as suggested by the expression 'moral certainty' that is sometimes used to paraphrase 'beyond a reasonable doubt.' Historically, the criminal standard used to be a device of moral comfort; it was meant to minimize the regret, remorse, and fear of God's punishment that jurors could experience in convicting a possibly innocent defendant (Whitman, 2008). Jury instructions used to stress the importance of minimizing future regret. For example, Justice Lemuel Shaw's statement of the BARD standard in Commonwealth v. Webster (1890) reads:

> Recognize that you are dealing with a person's life and liberty, and decide he is guilty only if you are confident that you will not be nagged by doubts about the propriety of your decision. Make a decision you can live with. 59 Mass. (5 Cush.) 295, 320.

The account of the BARD standard I have offerred here is open to both epistemic and moral considerations. For one, it is an epistemic account in virtue of the role that high probability and evidence play in it. For another, my account is also open to view the BARD standard as a device of *moral integrity*. This is particularly apparent in the resiliency condition. Resiliency is satisfied whenever the prosecutor's case has resisted the challenges leveled by the prosecutor. This very much depends on what the prosecution and the defense do. It is the prosecutor's duty to offer a complete narrative, well-supported by the evidence, a narrative that can be subject to extensive scrutiny. On the other hand, it is the defense's duty to challenge the proposed narrative in the most effective way. Finally, it is the fact-finders' duty to assess the exchange between the prosecutor and the defense, and then determine whether the prosecutor's case has survived the defense's challenges or not. The criminal standard is not prefixed once and for all; it emerges in the live confrontation between prosecution and defense—a confrontation which the fact-finders are called to assess and from which they are expected to draw the conclusion of guilt or innocence.

Upon concluding this chapter, I want to emphasize that my account of BARD is not in tension with legal probabilism; rather, it is compatible with it. I've suggested that, instead of estimating the probability of guilt simpliciter, we are better off with estimating the probability of the incriminating narrative. I've also suggested that, when one is trying to become clear about the criminal standard, the resiliency of the incriminating narrative and the completeness of the evidence are important factors to take into consideration. Both of them are not reducible to mere guilt probability; they raise different questions. Yet, these different questions can still be posed in a probabilistic way—i.e. 'how likely is it that the incriminating narrative will resist challenges?' and 'how likely is it that some items of evidence are missing?'—and they can therefore receive a probabilistic treatment. I leave it for future work to say how this can be done.

CHAPTER 8

WHEN IS DNA EVIDENCE ALONE ENOUGH TO CONVICT?

DNA evidence is widely praised. Some twenty years ago, we could even hear triumphant declarations that DNA evidence was infallible.[1] These exaggerations have disappeared from the literature, although they might still persist in the public opinion. Exaggerations aside, DNA evidence is rightfully regarded as one of the most powerful forms of incriminating evidence.[2] My goal in this chapter is to articulate a principled answer to the question: *When is DNA evidence alone enough to convict?* In answering this question, the chapter is organized as follows. In section 8.1, I examine the strengths and weaknesses of DNA evidence, and in particular its weaknesses. In section 8.2, I compare DNA evidence to fingerprint and eyewitness evidence. In the remaining sections, I examine two different models for reasoning with DNA evidence: the linear inference

[1] For a list of testimonies by forensic experts attesting to the infallibility of DNA evidence, see (Koehler, 1993).

[2] For a quick introduction to DNA evidence and its uses in the courtroom, see (Wasserman, 2008). For a more in-depth treatment, see (Kaye and Sensabaugh, 2000).

model and the narrative-based model. My conclusion is that DNA evidence is enough to justify a conviction provided it can support a well specified incriminating narrative which the defense had an effective opportunity to challenge.

8.1 FOUR REASONS FOR CAUTION

When traces of blood, semen, saliva, skin tissue, etc. are found at the crime scene, laboratory analyses can create a DNA *profile* from the traces. A DNA profile is a codified representation of select portions of the human genome, specifically, of those portions which tend to be different across individuals.[3] Once a profile is created, it is compared against the blood, semen, saliva, skin tissue, etc. of a suspect. The purpose of the comparison is to find a genetic *match*.[4] If a match is found, this would constitute strong

[3] The genetic material is mostly identical across different individuals, but small portions of it are highly variable. A DNA profile is a codified representation of the highly variable portions. More precisely, on a chromosome we can individuate specific positions, called *loci*. Each locus is "occupied" or "filled" by a particular DNA sequence, called an *allele*. In the eighties a British geneticist, Alec Jeffreys, discovered a number of loci whose alleles tend to vary across individuals. These loci are called Variable Number Tandem Repeats (VNTR) because what varies is the number of repetitions of a patter of nucleotides. We can visualize a pattern of nucleotides as a sequence of letters, such as ACA, so that, in a particular locus L1, the pattern ACA repeats itself a different number of times depending on the individual. Each VNTR locus is associated with the same repeating pattern for all individuals—e.g. L1 is associated with pattern ACA—but the number of repetitions of the pattern in the locus tend to vary across individuals. So, relative to some VNTR loci of interest, a DNA profile conveys information about the DNA sequences (alleles) at each locus, and in particular, it conveys information about the number of pattern repetitions at each locus. Importantly, the number of pattern repetitions is not measured by counting; this is impossible. What can be measured is the length of the VNRT locus. This means that, in fact, a DNA profile conveys information about the length of the VNTR loci of interest. The number of loci to be considered varies depending on the country. E.g., in the United States, the Combined DNA Index System (CODIS) created by the FBI requires that a DNA profile consists of 13 select loci. These 13 loci constitute a particular type of VNTR loci and they are called Short Tandem Repeats (STR) because the repeating pattern in the locus consists of a limited number of nucleotides. For more information, see Kaye (2010b).

[4] What counts as a match is not uncontroversial, for DNA profiles created from different samples are never identical; they are *more* or *less* congruent. Typically, a match between two profiles is declared when the two profiles are sufficiently similar, according to a fixed tolerance interval. Kaye (1993) proposes to replace the misleading language of 'match' and 'non-match' with statements about the profiles' degree of congruence. Besides, forensic experts might be mistaken when they declare a match (e.g. because of contamination, switching of the samples, or because the two DNA profiles do not match at all). On how laboratory errors affect the probative values of DNA evidence, see (Thompson et al., 2003).

but not infallible evidence that the suspect is the source of the traces found at the crime scene. What makes DNA evidence particularly powerful is that DNA profiles, albeit not unique, are highly discriminating because they occur very rarely.[5] The rarity of a DNA profile is expressed by a frequency, sometimes as astronomically small as 1 in 50 billion, representing the profile's expected frequency in a population.[6] The lower the frequency, the more discriminating the profile and consequently the more probative the match.

In the scholarly literature on DNA evidence, at least four reasons for exercising caution have been pointed out. *First*, DNA evidence is not directly probative of the ultimate issue of guilt, but only of intermediate propositions such as the identity of the person who left a DNA trace on the crime scene. *Second*, despite the fact that DNA evidence is very individual-specific, it cannot single out individuals uniquely. All we have is a statistical estimate of a DNA profile's expected frequency in a reference population; no claim of uniqueness is, in principle, warranted. *Third*, several commentators have argued that even a small chance of laboratory error renders DNA evidence significantly less probative. *Fourth*, we should be aware that a DNA profile's frequency is an extrapolation from a genetic model that has limited applicability. Let us examine each point more closely.

[5] Importantly, DNA profiles are not unique and two individuals might share the same profile; see (Weir, 2007) and (Saks and Koehler, 2008).

[6] A DNA profile, as explained in footnote 3, coveys information about the DNA sequences (alleles) at select loci. Each allele has a certain frequency among a given population and its expected frequency is calculated by counting how many times a certain allele shows up in a database of DNA profiles. The bigger the database, the better the frequency estimate. The FBI currently has a very extensive database, called CODIS, which contains millions of DNA profiles. Further, the frequency of the profile as a whole is computed by multiplying the frequencies of the single alleles. Consequently, the more alleles are included in the profile the lower the frequency. The multiplication of allele frequencies is justified on the assumption that the frequency of each allele is independent from that of the others. This assumption is not uncontroversial: scientists debated it extensively, though now it is generally accepted. Kaye (2010b) reconstructs the debate on the independence assumption. Buckleton (2005b) reminds us that astronomically low frequencies are based on statistical models which might not be valid when too many alleles are considered: the problem is that there is no empirical test to verify the statistical estimates concerning astronomically low frequencies.

8.1.1 Source and guilt

In any criminal case a number of factual propositions are to be decided, either intermediate or ultimate. The ultimate question is that of guilt, which is typically divided into two sub-questions, one concerning the identity of the individual who caused the harm (*actus reus*), and the other concerning the intentionality of the act (*mes rea*).[7] DNA evidence cannot address the question of intentionality, but it can play a role in answering the question of identity. Even with respect to the identity question, DNA evidence can, at its best, answer *the intermediate question of whether the accused is the person whose DNA traces were found at the crime scene*. In other words, even if a match between the DNA of the accused and the DNA traces could allow us to conclude—with absolute certainty—that the traces belong to the accused, this would only give us a compelling ground to answer 'Yes' to the following question:

(Q1) Accused = one whose DNA traces were found at the crime scene ?

Other questions cannot receive an explicitly positive answer by means of DNA evidence alone.[8] These questions are:

(Q2) Accused = one who left DNA traces at the crime scene?
(Q3) Accused = one who was involved in the crime?
(Q4) Accused = one who committed the harmful act?
(Q5) Accused = the guilty party?

Between (Q1) and (Q2) there is a leap, so to speak, because the DNA traces could have been purposely fabricated and left by anyone interested in misleading the investigators, without the accused having visited the crime scene at any point. Between (Q2) and (Q3) there is another leap, because having left a DNA trace indicates one's presence at the crimine scene at some point in time, yet it does not entail that one participated in the

[7] See, for instance, *California Penal Code*, sections 187-199.

[8] I am making essentially the same point as Koehler (1996a). He distinguishes three propositions (p. 865): one concerns guilt; another concerns whether the defendant had contact with the crime scene; the third concerns whether the defendant was the source of the DNA trace. DNA evidence bears on the third proposition only, and merely indirectly on the other two.

criminal act. In fact, one could have left a DNA trace without being involved in the crime, by virtue of visiting the crime scene before or after the crime. Between (Q3) and (Q4) there is still another leap, because participating in the crime does not entail being the main perpetrator. And finally, the leap beween (Q4) and (Q5) has to do with intentionality.[9]

How can we bridge the gap between (Q1) and (Q5)? Prosecutors must separately argue that the traces could not have been left innocently, so that whoever left them must be guilty; DNA evidence can complete this argument and show that the suspect *is* the source of the traces. The important point to note here is that a convincing story should be offerred of how the defendant left certain traces, not innocently, but while committing the crime. I will return to this point in section 8.4.

8.1.2 Lack of uniqueness

Defense lawyers sometimes advance what we might call *uniqueness challenges* against DNA evidence. They emphasize the indisputable fact that while DNA profiles occur rarely in a population, they are not unique.[10] Accordingly, DNA evidence would only be able to place the defendant or the accused among a group of people sharing the same DNA profile, without being able to single out one individual uniquely. The rationale behind these challenges is simple: If a DNA profile's estimated frequency is 1 in 10 million, and there are 30 million suspects, a match between the DNA from the crime scene and the DNA of the accused would only place the accused among a group of 3 people.

While these challenges report a well-known fact—i.e. DNA profiles occur rarely but need not be unique—they hint at a confused interpretation of DNA evidence. I think

[9] In probabilistic terms, the probability that the correct answer to question (Q1) is Yes, conditional on DNA evidence, could be quite high. Yet, this probability must decrease progressively as we consider questions (Q2) through (Q5). Even if it were unassailably true that the accused left a DNA trace, the probability that he committed the crime, must be below one, absent some other information; and this probability must be even lower if we consider the ultimate question of guilt.

[10] This has led at least one scientist to suggest that it would be better to carry out the necessary research to arrive at real DNA fingerprints—i.e., DNA profiles that are unique to individuals—rather than debating on the statistical subtleties underlying uniqueness claims. See Lewontin (1994), p. 261-262.

that a more adequate way to appreciate the uncertainty associated with the fact that DNA profiles are not unique is by means of Bayes' theorem, as illustrated in chapter 5. Focusing on the question of uniqueness is misleading because it leads to apply the wrong heuristics. Let me give an illustration of this.

In a California 1994 rape case involving a victim who could not identify the defendant, the only evidence linking defendant Frank Lee Soto to the crime scene was DNA evidence.[11] Besides the declaration of a match, the prosecutor experts testified as to various frequencies of the DNA profile in question, the highest of those being 1 in 189 million for the Hispanic population and 1 in 38 million for the Caucasian population. The defense attorney argued that DNA evidence 'should be treated like identification techniques which merely place the defendant within a *class* of possible suspects,' and he added that DNA evidence should be corroborated by some other evidence connecting the *specific* defendant to the crime scene before a conviction could be sustained.[12] In other words, the defense argument can be broken down, as follows:

> Given that (a) the proffered DNA evidence only placed the defendant among a group of suspects, however small that group may be, it follows that (b) the DNA evidence alone should not be enough to satisfy the standard of proof in a criminal case.

The inference from (a) to (b) is, I think, uncontroversial: if the evidence points to a pool of suspects that consists of at least two people, without making any distinction between them, each of them is equally likely to be the perpetrator, and thus the 'beyond any reasonable doubt' standard is not met. The appellate court rejected the defense argument, but it did not reject the inference from (a) to (b); it only rejected premise (a). While the defense asserted that someone else in California could share the same DNA profile, the court was persuaded that the experts testified to the contrary, saying that there was a 1 in 189 million frequency of the profile in the Hispanic population, with the population

[11] See *People v. Soto*, 48 Cal. App. 4th 924 (1994).

[12] See *Soto* 48 Cal. App. 4th at 946. The defendant required that the jury be assisted by the following instruction, reported in footnote 26 of the opinion: 'the analysis of the DNA in this case in and of itself does not identify the defendant as the perpetrator in this case. This evidence, in and of itself, is inadequate in absence of other corroboration, to identify the defendant as the perpetrator of the offense charged.'

of California totaling only 30 million.[13] The court, in other words, was persuaded that the DNA evidence alone pointed to a unique individual, the defendant. In other cases, similar uniqueness challenges have been raised and uniformly dismissed by appellate courts.[14] I should note, incidentally, that least one legal scholar has formulated similar challenges, requiring that DNA evidence be accompanied by more individualizing evidence to be enough to support a criminal conviction.[15]

Wrong heuristics. I think that a focus on uniqueness is misplaced. The reason is that, as appellate courts are asked to answer uniqueness challenges against DNA evidence, they are led to endorse the wrong heuristics. Appellate courts rely on the heuristics that *if* the reference population against which the DNA frequency is expressed exceeds in number the population of possible suspects, or the total earth population, *then* uniqueness claims are warranted. This heuristics is used quite extensively. It was exploited in

[13] See footnote 27 in the opinion.

[14] In *New York v. Rush*, 650 N. Y. S. 2d 631, 632 (Sup. Ct. 1995), the testimony of an expert is reported as saying that DNA comparison cannot result in absolute identification, unlike fingerprints. The appellate court did not find this point particularly important and concluded that 'the statistical probability that anyone else was the source of that semen are in 1 in 500 million is legally sufficient to support a guilty verdict' (*Rush*, 650 N. Y. S. 2d at 634). Incidentally, note that uniqueness claims in the case of fingerprints, although they are virtually unchallenged in the courts, are scientifically unwarranted. See Zabell (2005). In a later case, *Roberson v. Texas*, 16 S. W. 3d 156 (Tx. Ct. App. 2000), defendant argued that DNA evidence could only exclude an individual as a possible perpetrator, but could not be enough, by itself, to prove identity (*Roberson*, 16 S. W. 3d at 159). The court disagreed, essentially endorsing the *Rush* decision (*Roberson*, 16 S. W. 3d at 169). Finally, in a 2008 capital murder case, defendant raised uniqueness challenges and the court responded that a frequency as low of one in a quintillion provided enough specificity for arriving at the identity of the perpetrator. In *Missouri v. Abdelmalik*, 273 S.W. 3d, 61, 66 (Mo. Ct. App. 2008), the court's final ruling was as follows: 'We conclude that where, as here, DNA material is found in a location, quantity, and type inconsistent with casual contact and there is one in one quintillion likelihood that some else was the source of the material, the evidence is legally sufficient to support a guilty verdict.' Incidentally, note that the appellate courts in all three cases above committed the *prosecutor's fallacy*; see chapter 5.

[15] In commenting on a decision in a rape case concerning issues of paternity, *State v. Skipper*, 637, A.2d 1101 (Conn. Sup. Ct. 1994), Stein (2005) writes: 'because Skipper and the aborted fetus are far from being the only carriers of this genetic pattern, the weight of this probability does not satisfy the 'beyond any reasonable doubt standard' (p. 87). The *Skipper* case, however, is anomalous for a couple of reasons. First, the frequency of the DNA profile is extremely high; second, the issue to be determined is that of paternity. It is unclear what Stein would say if the DNA profile frequency was much lower. Concerning the determination of paternity, see (Kaye, 1989).

the *Soto* decision, noting that the population of California was roughly 30 million while the frequency of the DNA profile in question was in 1 in 189 million (among Hispanics). The discrepancy between 30 and 189 million was the key element to support the uniqueness claim, according to the Court in *Soto*.[16] Along similar lines, consider this exchange between an expert witness and a prosecutor:

> *Expert witness*: The final number was that you would expect to find only one individual in 234 billion that would have the same banding pattern [i.e., DNA profile] that we found in this case.
>
> *Prosector*: What is the total earth population, if you know?
>
> *EW*: Five billion.
>
> *P*: This is in excess of the number of people today.
>
> *EW*: Yes. Basically that's what that number ultimately means is that that pattern is unique within the population of this planet.[17]

And finally, here is one of the arguments against O. J. Simpson offered by the prosecutor:

> Ladies and gentlemen, his blood on the rear gate with that match, that makes him one in 57 billion people that could have left the blood . . . there is only five billion people on the planet. Ladies and Gentleman, that is an identification, okay, that proves it is his blood. Nobody else's on the planet; no one.[18]

As some commentators have noted, the simple heuristic for making inferences as to the uniqueness of a DNA profile is not, in general, statistically correct.[19] Even if the reference population against which the frequency is expressed exceeds the population

[16] Footnote 27 in the opinion reads: 'Counsel asserts the experts concluded in their probability calculation DNA type *could* have been held by someone else in California or the world. He submits no authority for this remark. The experts essentially testified to the contrary ("once in 189 million Hispanics"), assuming that the total population of California is 30.4 million, of which we take judicial notice' (*Soto*, 48 Cal. App. 4th at 947).

[17] See *Martinez v. State*, 549 So.2d 694 (Fla. App. 1989), cited in Kaye (1993), p. 118.

[18] *People v. Simpson*, Transcript (Superior Court, Los Angeles County, Closing argument by Ms. Clark), 1995 WL 672671 (September 26, 1995), cited in Saks and Koehler (2008), p. 203.

[19] See, among others, Koehler (1996a) and Balding (2005), chapter 9, using the expresion *uniqueness fallacy*.

of suspects, or the earth population, uniqueness claims cannot be made. This means that uniqueness challenges cannot be dismissed by simply pointing to the rarity of a DNA profile. The only acceptable answer to a uniqueness challenge is to say that, given a certain population model, a genetic profile is unique with x percent probability. But then, the question arises of how high the probability of uniqueness should be, and understandably, appellate courts are unlikely to answer this question in any direct way.

Where does this leaves us? Should we conclude that, therefore, DNA evidence cannot be enough for an identification unless the probability of uniqueness meets a reasonably high threshold? The problem is that a probability threshold is arbitrary and, when it is set too high, unduly restrictive. Uniqueness challenges cannot be easily dismissed, but at the same time, they set an unreasonably high bar on an identification. No evidence can uniquely identify the perpetrator—as we shall see, eyewitness evidence cannot; fingerprint evidence cannot—and thus expecting the evidence to do so would amount to a generalized challenge against any type of evidence.

Scholars are divided on the uniqueness issue. For instance, Saks and Koehler (2008) and Koehler and Saks (2010) hold that uniqueness claims cannot be made, yet Kaye (2010a, 2013) holds that—with qualifications—they can be made.[20] Other scholars, e.g. Balding (2005), believe that the uniqueness format for presenting DNA evidence is too restrictive and suggest to adopt a Bayesian format. I agree. (See chapter 5 in which I contrast the Bayesian approach and the uniqueness approach.) The reason for the superiority of the Bayesian approach is roughly that it can incorporate more sources of uncertainty besides a genetic profile's possible lack of uniqueness. What are, then, the other sources of uncertainty?

Population of suspects. A judgment of uniqueness in a courtroom will always be relative to a population of suspects; it will never be an absolute judgment. Criminal

[20] Koehler and Saks (2010) talk about an *individualization fallacy* when unwarranted claims about the uniqueness of a DNA profile are made. They forcefully argue that 'no sund and rigourous evidence supports the assumption of unique individualization.' The claim is contested. In fact, Kaye (2010a) is less categorical and holds that 'a scientifically defensible opinion as to individualization is still attainable in some situations.'

trials, after all, are not interested in establishing the metaphysical uniqueness of a given genetic profile. For the sake of illustration, suppose the population of suspects has been identified by means of an eyewitness testifying e.g. that the perpetrator was a male adult in his mid-thirties. Given the information provided by the eyewitness, one could make the guess that males in the their mid thirties are, say, 20 million, so that the population of suspects will be narrowed down to 20 million people only. But how is the number '20 million' arrived at? Have reliable statistical data been consulted? Probably not. In other cases, the police might have selected a very limited number of suspects because of some hints and clues received by various informants. But are the few suspects the only possible suspects, thus completely ruling out everyone else? Finally, cases in which the suspects are so few in number that a laboratory comparison is possible between each suspect's DNA and the crime scene DNA would seem to put an end to the dispute about uniqueness: if only one among the few suspects match, then—one could reason—he must be the only one that DNA evidence links to the crime scene. This would be correct, but only provided that among the few suspects singled out by the police there is certainly the one who left the DNA trace on the crime scene; and this need not be at all certain.

In general, the problem is that the population of suspects might have been identified in a questionable way; the fact that one individual is not included in it does not entail that he is completely ruled out. Identifying a fixed and rigidly determined population of suspects suggests a dichotomy between those that are certainly "in" and those that are certainly "out." This dichotomy is fictitious—one could say—because people are simply "more" or "less" likely to be the source of the crime scene DNA.

Uniqueness and laboratory errors. Besides difficulties in identifying the population of suspects, another source of uncertainty is constituted by laboratory errors, which are strongly linked with the question of uniqueness. To see why, observe that the *legal* question of uniqueness can be phrased as follows: (u) whether no one other than the defendant has the DNA profile in question, *given that the defendant has it*; or more carefully, $(u*)$ whether no one other than the defendant has the DNA profile in question, *given that the defendant is shown to match*. The difference between (u) and $(u*)$ is due the fact that even if the defendant's DNA matches the one on the crime scene, the match could be wrong, and so the match does not entail that the defendant actually has the DNA profile in question. Yet, if we set aside errors related with the declaration of a match, and assume that the declaration of a match entails that the defendant has the profile in question, then (u) and $(u*)$ boil down to the same question. Laboratory errors, however, are not negligible, and to them I now turn.

8.1.3 Laboratory errors

Laboratory errors have occurred and will occur because DNA forensic technology involves a significant part of human intervention, consisting of collecting the DNA samples, storing them, labelling them, analysing them, interpreting the results, etc. Laboratory errors occur and they affect the probative value of DNA evidence. This is a truism, but whether or not it should be a reason for concern is not uncontroversial. In fact, the 1992 report on DNA evidence by the National Research Council warned against the dangers associated with laboratory errors and how these can affect the probative value of DNA evidence in significant ways (NRC, 1992). In contrast, the subsequent 1996 report by the National Research Council downplayed the need to be concerned with laboratory errors (NRC, 1996). But as observed in chapter 5, even a relatively small laboratory error rate can significantly diminish the probative value of DNA evidence, by dramatically reducing the likelihood ratio.[21] The impact of laboratory error rate is an indication that no matter how low a profile's frequency turns out to be, it will be of relatively little significance if the frequency is unaccompanied by an estimate of the laboratory error rates (Thompson et al., 2003).

8.1.4 Probability models and frequencies

If DNA profiles cannot be said to be unique, we should also be careful to understand the meaning of the incredibly low frequencies of DNA profiles. What does it mean that a DNA profile has a frequency of one in 60 billion people? It is wrong to think that we

[21] For a more systematic discussion of this point, see Thompson et al. (2003).

are dealing with actual frequencies, as though in a population of 60 billion people there would be one DNA profile of a certain type. Who are the 60 billion people we talking about? We should not forget that DNA profiles' frequencies are the result of projections yielded by genetic and statistical models, so it is more appropriate to talk about estimated frequencies or probabilities. Now, whenever frequencies as low as in 1 in 60 billion or lower are introduced in court, some defense lawyers will raise the plausible objection that such astronomically small numbers are beyond the comprehension of jurors and lay people—those numbers are devoid of any meaning. While these comments could be marked as mere rhetoric, they do posses some bite. Let me explain.

The frequencies of DNA profiles are calculated using probabilistic and genetic models of how genetic features are distributed among individuals. The precise details of how these frequencies are calculated are not essential here, but it is important to emphasize that *these frequencies are an extrapolation from population genetic models*.[22] There is no easy way to empirically test the correctness of those frequencies, for how would one empirically test that the frequency of a DNA profile is 1 in 1 billion? To be sure, we can rely on those frequencies because we can rely on the models from which they are derived: as long as the models are valid, the derived frequencies can be relied upon. The trouble is that the more extreme the frequencies, the more extreme the extrapolation from those models; and when the extrapolation reaches its extreme, it is increasingly uncertain whether the models are still valid. Along similar lines, one commentator writes:

> . . . we recognize that we are considering an extreme extrapolation using these models. We are not operating near the center of their prediction range where they are more testable and tested. The models have been extensively tested in this central range and there is some considerable reason to believe

[22] Here is how a forensic scholar, Ian Evett, quoted in Buckleton (2005b), section 3.1, describes the process:

> In the DNA context, I take some numbers (that are estimates of things like allele proportions . . .) and stick them into a formula. Out comes a number and on the basis of that I assign ... a probability [*or a frequency, in the terminology I have been using*]. That is a personal, subjective probability, which incorporates a set of beliefs with regard to the reliability/robustness of the underlying model.

> that they are robust there, but they are still models and the probabilities produced by them are still untestable. (Buckleton, 2005b, section 3.4.5.).

These observations are relevant for the issue of uniqueness. If a profile's frequency is high, no uniqueness claim would be warranted (at least, against the total earth population). But even if the frequency is incredibly low, the uniqueness claim would still be problematic because the genetic models from which the low frequency is extrapolated might not be 'operating near the center of their prediction range.' The lesson to be learned here, I think, is not that DNA evidence is worthless, but that an exclusive focus on uniqueness and on how many people have such-and-such DNA profile is misleading. The question of uniqueness is important, but it is not the only question to ask while evaluating DNA evidence. For instance, if the population of suspects is small enough and has been identified though other weighty evidence, DNA evidence is very useful in establishing an identification. This suggests that police and prosecutors should not rely on DNA evidence alone as incriminating evidence, but should collect as much as other evidence as possible. DNA evidence is far from a truth-machine and the old-fashioned ways of doing investigations, so to speak, are not dispensable.

8.2 Comparisons

I have examined some of the limitations underlying DNA evidence in criminal trials. We should keep things in perspective, however. These limitations do not automatically make DNA evidence a weak form of evidence. To acquire a more balanced view, it is instructive to compare DNA evidence to other important forms of criminal evidence: fingerprint and eyewitness evidence.

8.2.1 Fingerprint evidence

Fingerprint evidence has long been regarded as the gold standard of criminal evidence, and some decades ago it was not uncommon to read that fingerprint evidence could provide an infallible and absolute identification.[23] Presumably, such blind faith in fingerprint evidence was based on two assumptions: that forensic experts commit few mistakes, or none at all; and that fingerprints are unique to individuals, so that if the fingerprints found at the crime scene match with the defendant's fingerprints, the two fingerprint samples must come from the same individual. Both assumptions, though, are unwarranted, and the scholarly literature is now well aware of this fact (Kaye, 2003; Zabell, 2005; Dror et al., 2006; Thompson and Cole, 2007). Nonetheless, we can still hear expert witnesses testifying to the uniqueness of fingerprints, and the FBI in a recent report insisted that fingerprints are unique to individuals.[24]

While the shortcomings of fingerprint evidence are relatively well known, the scholarly literature has paid little attention to the question of its probative value in comparison to DNA evidence.[25] Fingerprint evidence works in ways that are similar to DNA evidence. In the commission of criminal acts, perpetrators sometimes leave fingerprint traces on the weapon used, at the crime scene or in the surroundings, or on the victim's body. When the investigators recover fingerprint traces, they will compare them with those of a suspect. If they turn out to match, this will constitute evidence against the suspect, although a match does not yet show that the suspect is guilty. Leaving fingerprint traces, after all, is still compatible with innocent conduct.

The reasons for caution about fingerprint evidence are similar to the four reasons I've listed while discussing DNA evidence. If you recall, these four reasons were: inability to address the ultimate issue of guilt; lack of uniqueness; testing errors; contested trustworthiness of the astronomically low frequencies. Let us see how these four items apply to fingerprint evidence. First, the ultimate issue of guilt cannot be immediately proven from fingerprint traces alone; the investigators should offer a compelling reconstruction of the suspect's involvement in the crime. Second, fingerprint profiles are not

[23] E.g. Jamison v. State, 354 S.W.2d 252, 255 (Tenn. 1962) ("all the fingerprints that have ever been taken as far as reported cases are concerned run into an infinite number and no two have ever been found alike. The courts generally have reached the conclusion that the evidence of this kind is infallible because of its conclusiveness").

[24] The study was sponsored by the FBI and was conducted by Meagher, Budowle, and Ziesig in 1999. For a critical discussion of the FBI study, see e.g. (Stoney, 2001; Kaye, 2003).

[25] But see (Zabell, 2005).

unique. This might appear surprising to some, but as a matter of fact no one has even demonstrated that fingerprints are unique (Zabell, 2005). Third, expert testimonies that a match exists between two prints turn out to be extremely subjective and variable across different experts. Alarmingly, experts can disagree with each other (Dror et al., 2006). Finally, the fourth problem is even more damning than in the case of DNA evidence: we do not have statistical estimates about fingerprints at all, let alone dependable ones, yet investigators often assume that prints are unique. All in all, the reliability of fingerprint evidence is currently under severe scrutiny by the academic community (NRC, 2009).

8.2.2 Eyewitness identification

When it comes to eyewitness evidence, the current intellectual trend among scholars and informed judges is well reflected in the following statement by the highest court of the state of New York:

> ...the perils of eyewitness identification testimony far exceed those presented by DNA expert testimony. Where the prosecutor is confronted with an irreconcilable conflict between eyewitness identification evidence and DNA identification evidence, it is likely to rely on the DNA evidence. People v. Rush (1995), 630 N.Y.S.2d 631, 634.[26]

The quotation does not express a position that is unanimously shared by judges and scholars, for some have warned against the perils of DNA evidence, especially insofar as its probabilistic and statistical underpinnings are concerned.[27] I, too, am inclined to suggest some caution. I also think that proclaiming the superiority of DNA evidence to eyewitness testimony is misleading. To be sure, eyewitness evidence is not in high regard today—and justifiably so. As the literature on the subject grows, judges and lay

[26] The *Rush* decision was followed by other appellate courts across the United States. See e.g. U.S. v. Wright, 215 F.3d 1020 (9th Cir. 200); Roberson v. State, 16 S.W.3d 156 (Tex. App. 2000); State v. Abdelmalik, 273 S.W.3d 61 (Mo. App. 2008). For a recent argument that DNA evidence alone can be sufficient to sustain a criminal conviction, see (Roth, 2010).

[27] See e.g. (Stein, 2005, p. 87) (arguing that DNA evidence should be accompanied by non-statistical and case-specific evidence); (Lingertwood, 2011) (arguing that proof beyond a reasonable doubt is inductive and not probabilistic).

people are becoming increasingly aware of the shortcomings associated with memory and perception.[28] Despite these limitations, I think that DNA evidence and eyewitness testimony have their own strengths and weaknesses, so that, on an absolute scale, neither of the two is evidentially stronger (or better) than the other. The strength of DNA evidence is that it rests on reliable statistical estimates, whereas eyewitness testimony rests on the hunch of a witness. On the other hand, the strength of eyewitness testimony is that it can address the ultimate issue of a case, whereas DNA evidence can only address intermediate issues.

In describing eyewitness evidence, I shall assume a very simple and abstract model of how eyewitness identification works. The model is divided in three stages:

> PERCEPTION STAGE. First, an eyewitness sees the culprit at the crime scene. With varying degrees of precision, the witness sees what the culprit looks like and what the culprit is doing.
>
> MNEMONIC STAGE. Next, the witness retains in his memory what he saw. That is, he retains a number of facial and physical features possessed by the culprit (e.g. tall, blue eyes, pale, blond, young), or more holistically, he has memory of the entire face without breaking it down into separable features.[29] The witness also retains in his memory a sequence of actions which the culprit performed.
>
> IDENTIFICATION AND RECOUNTING STAGE. Finally, when the eyewitness is asked to identify the culprit among a number of suspects, he will look for a *matching individual*, i.e. an individual who has the features which the witness remembers the culprit to have. And also, when the eyewitness is asked to recount what happened, he will recount the sequence of actions which he remembers the culprit to have performed.

[28] This increased awareness is also a product of the public visibility of *The Innocence Project* (www.innocenceproject.org). More than three hundred cases of wrongful convictions have been documented; in seventy percent of them eyewitness misidentification turned out to be responsible for the wrongful conviction. Recently, the Supreme Court of New Jersey in State v. Henderson (2011) has commented on the shortcoming of eyewitness identification and has suggested more stringent standards of admissibility.

[29] Scientific evidence seems to suggest that we recognize faces holistically, for our performance in recalling individual features is worse than our performance in recalling entire faces (Tanaka and Farah, 1993). Orientation is also important: we are better at recognizing upright faces rather than inverted faces, a phenomenon called the *Thatcher illusion* (Thomson, 1980).

Similarities with DNA evidence. Eyewitness identification and DNA evidence work in similar ways. In both cases, the process of identification relies—implicitly or explicitly—on a match. The features that are relevant for a match can be phenotypical, as in the case of eyewitness identification, or genotypical, as in the case of DNA evidence identification. To strengthen the parallelism, it is useful to consider the different ways DNA and eyewitness evidence could go wrong.

An eyewitness identification can go wrong in at least three ways (leaving aside cases of fabrication and purposefully false testimony). These parallel three ways DNA evidence identifications can go wrong. Here is an illustration:

> (*First*) The culprit has features F, but the eyewitness believes that the culprit has features F^*, because of misperception, mnemonic confusion, etc. Similarly, the DNA sample at the crime scene is contaminated or becomes contaminated because of laboratory mishandling. Contamination and laboratory mishandling are very much like misperception and memory failures for an eyewitness.
>
> (*Second*) The culprit has features F and the eyewitness correctly believes that the culprit has features F, yet at the time of the identification, he wrongly identifies as the culprit a suspect who has features F^*. Similarly, the reported match between the DNA sample from the crime scene and the defendant's DNA does not correspond to an actual match.
>
> (*Third*) The culprit has features F, the eyewitness correctly believes the culprit has them and correctly identifies a suspect who has features F as having them; the problem is that two individuals share features F, and the eyewitness happens to incriminate the innocent one. Similarly, more than one individual has the DNA profile in question.

In the case of eyewitness evidence, the first type of error is well-known. Even when subjects focus on a series of events, they often fail to notice important details. For instance, subjects whose attention is focused on watching people play basketball might not notice that a giant gorilla has made its appearance several times (Simons and Chabris, 1999). Further, a witness memory is easily prone to manipulation, and by asking leading questions (intentionally or otherwise), false facts can be easily inserted in a witness' recollection (Loftus, 1996).

Concerning the second type of error, confusions and mistakes that can arise during the identification are well known. Police lineups can be prejudicially suggestive, especially when the police exercise undue pressure on the witness to identify the culprit. To be sure, the second kind of error deserves a more careful formulation. As it is, the assumption is that, in both the DNA and the eyewitness case, there is a sharp difference between a match and a non-match. This is questionable. Whenever two DNA profiles are compared in a laboratory, there is nothing like a black-and-white match, or the lack thereof. As often happens in science, it is only a matter of degrees. Clearly, the two profiles are not identical, and there may be smaller or greater factors making them different. It would be more appropriate to talk about degrees of similarity or congruence between two DNA profiles, rather than a match or a non-match (Kaye, 1993).

The same consideration can apply to eyewitness identification. There are no sharp distinctions. All an eyewitness can do is to judge that an individual whom he saw at the crime scene is *more or less* similar to an individual whom he now sees at the police station or on a photograph. This is partly confirmed by the *relative judgment theory* of eyewitness identification (Wells et al., 2006). According to the theory, when an eyewitness is asked to identify the culprit from a group of suspects, he will point to the suspect who is the *most similar* to (his recollection of) the culprit. So, depending on the group of suspects from whom he has to choose, the witness might point to two different people as the culprit, both being the most similar to the cuprit *relative to the select group of suspects*.

Finally, the third kind of error, to my knowledge, is little discussed. It is often assumed that if the perception and the memory of the witness are good enough, and the lineup is conducted in the most careful way, eyewitness recognition is likely to be successful. But this depends on the rarity of the "matching features" which were used by the eyewitness—implicitly or explicitly—in the act of identification. Features that the eyewitness takes to be very rare might not be rare at all.

Dissimilarities from DNA evidence. Besides these similarities, eyewitness and DNA evidence also differ. They differ quantitatively. Though both a witness and a laboratory make mistakes, laboratory procedures should be more reliable than eyewitness identifications. This means that the rate of false positives and negatives should be significantly lower in DNA laboratory analyses than in eyewitness identifications.

There are also important qualitative differences. DNA evidence does not speak directly to the question of guilt. Recall the distinction I made between questions (Q1) through (Q5), repeated here for convenience:

(Q1) Accused = one whose DNA traces were found at the crime scene ?
(Q2) Accused = one who left DNA traces at the crime scene?
(Q3) Accused = one who was involved in the crime?
(Q4) Accused = one who committed the harmful act?
(Q5) Accused = the guilty party?

Eyewitness evidence can speak directly to questions (Q3) and (Q4), and partly to question (Q5). It can be more directly relevant to the propositions to be proven than DNA evidence; this is why DNA evidence counts as circumstantial evidence, while eyewitness testimony can count as direct evidence. For this reason, an incriminating case entirely based on DNA evidence might be weaker than a case entirely based on eyewitness evidence, unless the prosecutor can show that a positive answer to (Q1), based on DNA evidence, licences the presumption that the answer to (Q4) and (Q5) should be positive, as well. This can occur, for instance, when the prosecutor can show that the presence of the defendant's DNA at the crime scene is incompatible with causal and unintentional contact.

Finally, another difference is that, with DNA evidence, there are statistical estimates of a DNA profile's frequency. The estimates allow experts to have meaningful disagreements about the rarity of DNA profiles. Not so for eyewitness evidence. The phenotypical features used by the eyewitness for identifying the culprit are not explicitly known, and their frequencies across a population are not statistically known either. I would be surprised, however, if the frequencies of the phenotypical features were lower than the frequencies of some DNA profiles. If they are not lower, then the uniqueness challenges I discussed in section 8.1.2 can be levelled against eyewitness identification evidence as well. The only reasons why these challenges would not apply are: first, that the phenotypical features used for identification are unique; and second, that the process of eyewitness identification does not work through a "feature-matching" schema. I am inclined to think that both these possibilities are unlikely.[30]

All in all, the strength of DNA evidence is that it rests on reliable statistical estimates and rigourous laboratory testing procedures, whereas eyewitness testimony rests on the implicit, unspecified, and uncontrollable hunch of a witness. On the other hand, the strength of eyewitness testimony is that it can address the ultimate issue of a case, whereas DNA evidence can only address intermediate issues.

8.2.3 Is all evidence statistical?

After this comparison of DNA, fingerprint, and eyewitness evidence, we can ask whether they are all forms of statistical evidence. There is a sense in which they all are, but we should be careful here. The difference between DNA and fingerprint evidence is minimal, except that we have statistical estimates of the frequency of DNA profiles, while we do not have such estimates for fingerprint profiles. Can we say that both types of evidence are statistical? In fact, fingerprint evidence is not statistical because we have no statistical estimate of the frequency of fingerprint profiles. But it should be statistical because to properly assess the significance of a fingerprint match we do need statistics. When and if fingerprint evidence improves, it will be as statistical as DNA evidence.

Eyewitness evidence and DNA evidence differ significantly in that an eyewitness can reconstruct more facts about the crime than DNA evidence can. A witness can

[30] Psychological findings suggest that we can discriminate between different faces if more differences in facial features are apparent (Bradshaw and Wallance, 1971). Findings also suggest that the processing in face recognition is holistic rather than piecemeal, because performance in recalling individual features is worse than performance in recalling entire faces (Tanaka and Farah, 1993). Orientation seems also very important: we are better at recognizing upright faces rather than inverted faces, a phenomenon sometimes called the *Thatcher illusion* (Thomson, 1980). Other findings suggest that pigmentation, shading, and depth are important for face recognition, and that we perform poorly at recognizing one-dimensional face drawings (Davies et al., 1978). This cursory review shows that there is good evidence that face recognition is based on the recognition of features, whatever these may be. Whether facial features are unique is hard to say. To my knowledge, there has not been any study of the "rarity of faces," in the same way in which we have statistical estimates of genetic profiles frequencies.

tell what happened at the crime scene, what the culprit was doing etc. DNA evidence has a more limited say in this matter. The other difference is that DNA laboratory analyses should be more reliable than eyewitness identifications. On the other hand, DNA and eyewitness evidence both rely on a "feature matching procedure" to identify the culprit. DNA evidence relies on genetic profiles, while an eyewitness relies on his memory of what the culprit looked like; presumably a witness' memory consists in a set of features and traits belonging to the culprit, what we may call an *implicit, identifying description*. Can we say, then, that eyewitness evidence is statistical? We have no statistical estimates about the implicit descriptions which eyewitnesses use. We tacitly rely on the assumption that few people or only one person fit them. But if so, statistical estimates could help us properly assess the probative value of eyewitness identifications. Eyewitness identifications therefore would be enhanced if they were supported by statistical estimates. The problem is that it is hard to verbalize the features and traits which eyewitnesses use to carry out their identifications. All we can say for now is that eyewitness evidence is a "covert" type of statistical evidence, while DNA evidence is an "overt" type.

8.3 LINEAR INFERENCE MODEL

After a comparative analysis, I now examine DNA evidence and its probative value more closely. At its simplest, we can single out two essential ingredients in DNA evidence:

> MATCH. The declaration of a match by a DNA expert;
>
> FREQUENCY. A statistical estimate of the frequency of the DNA profile for which the match is declared.

In addition, DNA evidence can be accompanied by information about the forensic and laboratory procedures carried out to collect, keep, and analyze the DNA samples from the crime scene and from the suspect. It can also be accompanied by information about the shape and arrangement of the genetic material found at the crime scene.

In recent years, and in conjunction with the phenomenon of *cold-hit* cases—i.e. cases in which the suspect is identified through a database search returning a genetic match, a search in which several million DNA profiles are compared; see section 8.6—courts and scholars have confronted the question of whether a criminal conviction is acceptable on the sole basis of DNA evidence (Roth, 2010; Lingertwood, 2011). This question, however, is ambiguous between two readings. If the DNA evidence in question consists of MATCH and FREQUENCY only, the answer must surely be negative; if it also consists of the crime traces, the answer is open to debate. The reason should be clear enough by now. As suggested earlier, a genetic match, even when it is factually correct, is not probative of the ultimate issue of guilt. A genetic match is only probative of intermediate propositions, e.g. whether the suspect is the source of the crime traces. Arriving at the conclusion that the suspect is guilty, or at the conclusion that he was present or involved in the crime, requires something more.

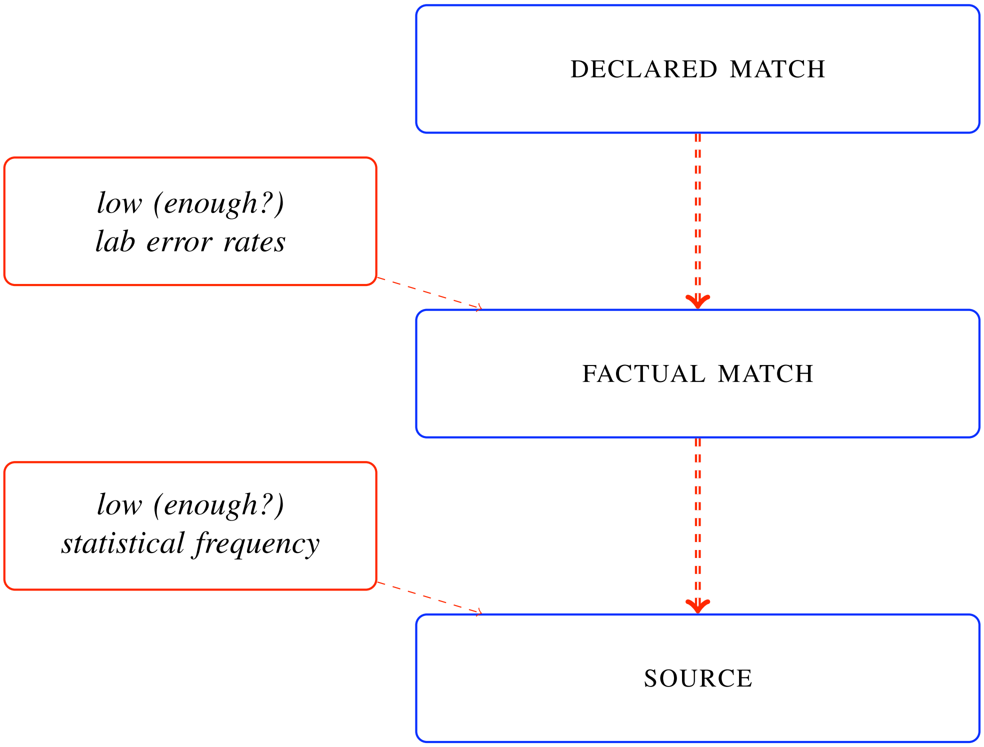
It is useful to adopt what I call a *linear inference model* and list the inferences one can draw from the finding of a genetic match:

> DECLARED MATCH: laboratory analyses declare a match;
>
> FACTUAL MATCH: the suspect's DNA profile and the crime traces, in fact, match;
>
> SOURCE: the suspect is the genetic source of the crime traces;
>
> PRESENCE: the suspect was present at the crime scene;
>
> ACTUS REUS: the suspect committed the harmful act;
>
> GUILT: the suspect committed the harmful act intentionally.

Let us examine each inference step more closely.

8.3.1 From declared match to source

When laboratory analyses declare the existence of a genetic match, this establishes that there is a factual match between the suspect and the crime traces only insofar as the laboratory analyses carry a relatively small margin of error. How small such a margin should be is a difficult question which I am going to set aside for the time being. Absent some information about the laboratory procedures and laboratory error rates, the inference from a declared match to a factual match would rest on shaky grounds.



The next inferential step is from the existence of a factual match to claiming that the suspect is the *source* of the genetic material found at the crime scene. This step is justified only insofar as the DNA profile for which the match is declared occurs rarely enough. Ideally, the profile should be unique, and if so, the inferential step from FACTUAL MATCH to SOURCE would be watertight. As seen earlier, DNA profiles cannot be proven to be unique, and at best we can expect them to be sufficiently rare. And again, how rare the profiles should be is a difficult question which I should set aside for the moment.

As the diagram shows, three items are required to establish source: a declared match; a statistical estimate of the frequency of the DNA profile; and information about the laboratory procedures and laboratory error rates.[31] Now, if DNA evidence consisted of

[31] Wigmore (1913) famously devised a chart method to analyze trial evidence; see also (Twining, 1986; Schum and Kadane, 1996; Goodwin, 2000; Anderson et al., 2005). My chart is a simplification of a

only these three items, it would only support the reasoning that leads up to SOURCE, and nothing beyond it. DNA evidence, however, can also include information about the shape and arrangement of the genetic material found at the crime scene, what we can call *trace evidence*. To be sure, it can be a point of terminological debate whether DNA evidence includes trace evidence. Let's say that, loosely speaking, DNA evidence does include it. After all, if genetic material is found at the crime scene, in some shape and arrangement, it also constitutes trace evidence in the traditional sense of the term.

A plausible position to take here is to say that DNA evidence is enough to establish guilt so long as one establishes that the presence of the genetic material at the crime scene is incompatible with innocent contact. In line with this approach, an Appellate Court wrote:

> We conclude that where, as here, DNA material is found in a location, quantity, and type inconsistent with casual contact and there is one in one quintillion likelihood that some else was the source of the material, the evidence is legally sufficient to support a guilty verdict. Missouri v. Abdelmalik, 273 S.W. 3d, 61, 66 (Mo. Ct. App. 2008).

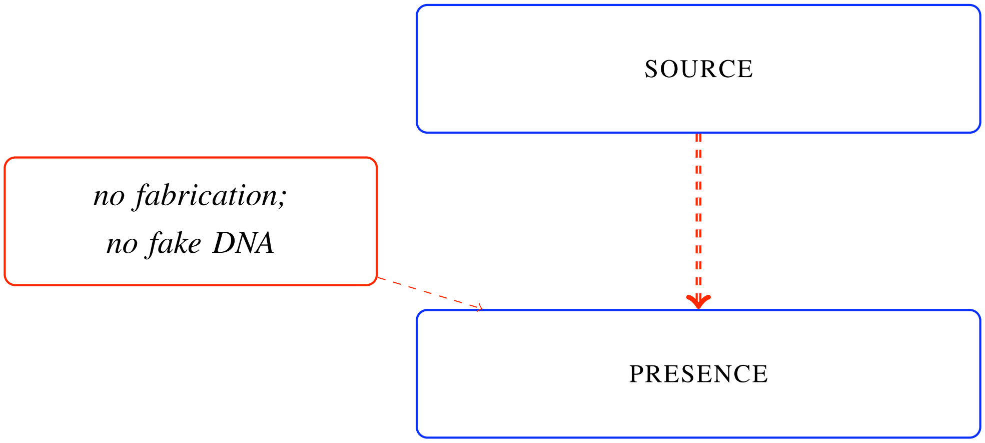
But what does it mean to say that 'DNA material . . . is inconsistent with casual contact'? According to my list of inferences above, the reasoning from SOURCE to GUILT goes through two intermediary propositions—PRESENCE and ACTUS REUS—although there might very well be additional intermediary propositions I've glossed over.

8.3.2 From source to presence

The step from SOURCE to PRESENCE is justified only insofar as one rules out the possibility that the DNA traces were fabricated to mislead the investigators, that is, only insofar as one establishes that the suspect actually visited the crime scene at some point. Investigators are well aware of the possibility of framing, but DNA evidence offers a more subtle opportunity for criminals to mislead the investigators. Studies have recently shown that it is in principle possible to synthesize and implant fake genetic material

Wigmore chart.

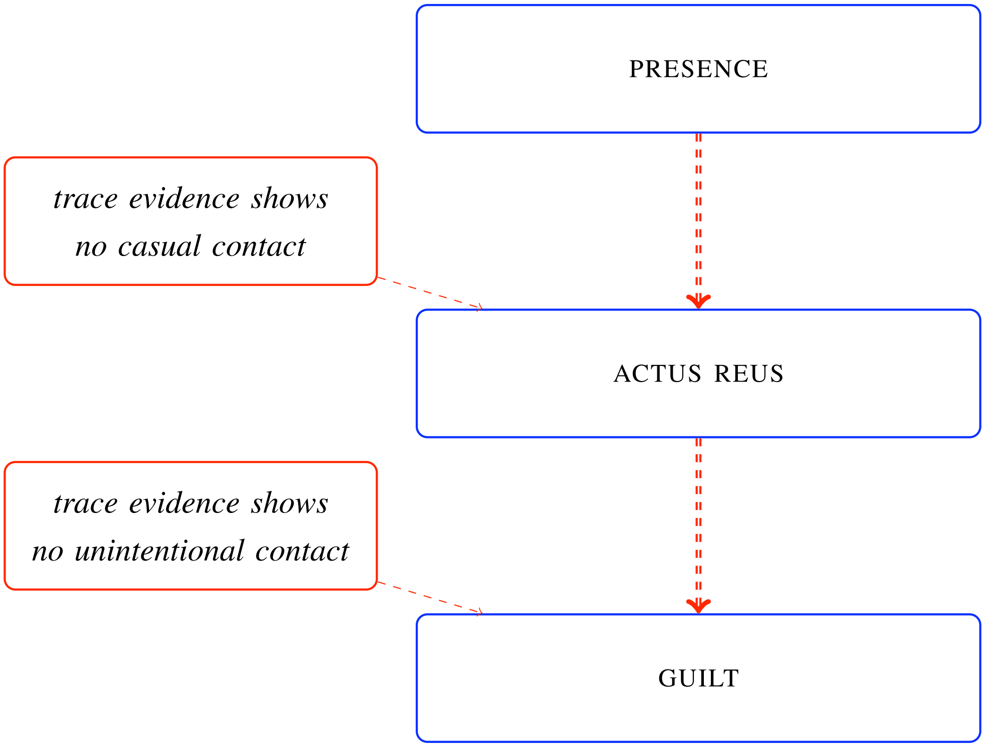
without modifying the crime scene in any visible way (Frumkin et al., 2009). If this possibility were to be heavily exploited by criminals in the future, the inference from source to presence would rest on shaky grounds, and the overall usefulness of DNA evidence in criminal trials would be severely compromised. What is particularly alarming is that standard DNA forensic technology cannot distinguish genuine genetic material from synthetic material. Suitably enhanced laboratory procedures, however, can distinguish fake from real genetic material.



8.3.3 *From presence to guilt*

Let us now turn to the inferential step from PRESENCE to ACTUS REUS and finally GUILT. Here, too, there is some uncertainty because the suspect could have left a trace at the crime scene without being involved in the crime; he might have visited the crime scene before or after the crime had been committed. The issue is further complicated by the fact that being involved, in some way or another, in the harmful act does not entail being the perpetrator of the act. Finally, in order to establish the ultimate question of guilt, one needs to establish the intentionality of the act, not only its commission.

The inference from PRESENCE to GUILT, however, might be less problematic than the inference from SOURCE to PRESENCE. The reason is that when DNA evidence includes information about the shape and arrangement of the crime traces, this information is useful to reconstruct how the crime occurred and whether the presence of the traces is incompatible with innocent contact (James et al., 2005). As I shall argue more extensively later, the inferential gap between presence and guilt can be filled provided a convincing narrative is offered of how the defendant left the genetic traces, not innocently, but while committing the crime.


PRESENCE
trace evidence shows no casual contact
ACTUS REUS
trace evidence shows no unintentional contact
GUILT

8.3.4 Three difficulties

My exposition so far has identified, albeit somewhat arbitrarily, three parts in the inferential chain leading to guilt: (1) from MATCH to SOURCE; (2) from SOURCE to PRESENCE; (3) from PRESENCE to GUILT. Let us examine each part more closely and the difficulties they pose for the linear inference model.

(1) The *up-to-source part* has been the focus of extensive scholarly attention. This part is naturally analyzed in probabilistic terms by using Bayes' theorem. This is the most natural way to proceed, given that both statistical frequencies and laboratory error

rates consist of quantitative information. The result is that the proposition 'source' will have attached a probability. (I have sketched how this probability can be estimated in chapter 5.) This probability estimate raises a difficult question: How high should the source probability be to meet the reasonable doubt standard? This a shorter way to ask two questions which I set aside earlier: How low should the profile's statistical frequency be? How low should the lab error rate be? Intuitively, the lower the error rate and the lower the frequency, the higher the source probability.

The question about the source probability can hardly receive an answer, unless we settle on some arbitrarily high value. But what would it be? The most one can say here is that this value should not be too low.[32] Alternatively, we might decide not to assign a probability to 'source' and proceed informally. This hardly solves the problem: it simply reformulates a probabilistic question—i.e. how high should the source probability be?—into a non-probabilistic one—i.e. how strongly supported should the proposition 'source' be? All in all, there is no easy answer within the linear inference model. Unsurprisingly, on source attribution and the role of experts, a well-regarded DNA forensic expert writes:

> My feeling is that we would be unwise to conclude the same source because it is not our place to do so. If we do so, . . . I would like us to make transparent that we have subjectively decided to round a probability ESTIMATE off to zero. On balance I cannot see much positive coming from a policy of declaring a common source. (Buckleton, 2005b, sec. 3.4.6)

(2) Next, we have the *up-to-presence part*, which raises another difficulty. Suppose it is very unlikely that a criminal would implant fake genetic material; if so, is the inference from SOURCE to PRESENCE justified? Or is such an inference still unwarranted in absence of evidence that rules out the possibility of DNA evidence fabrication? For one thing, one could argue that if one establishes SOURCE, a rebuttable presumption should

[32] Incidentally, note that criminal law does not require that source be established beyond a reasonable doubt. We might thus be tempted to dismiss the question altogether, but we should not forget that if source is weakly supported, then anything that rests on it will be weakly supported as well, e.g. *actus reus* or guilt itself. The question of how source is established beyond a reasonable doubt cannot be dismissed so quickly.

allow one to infer PRESENCE. In other words, so long as the defense does not present any evidence against the inference from SOURCE to PRESENCE, one could argue the inference must be presumed to be acceptable. On the other hand, one could also argue that allowing such a rebuttable presumption violates the presumption of innocence, for it is not the task of the defense to disprove that the defendant was present at the crime scene. We are in a dilemma here.

(3) Finally, the *up-to-guilt part* is what makes DNA evidence similar to other forms of trace evidence, e.g. fingerprints, glass evidence, blood evidence. Here the statistical and probabilistic underpinnings of DNA are out of the picture. What matters is to have in sight a reconstruction of what happened and of how the traces were formed as a consequence of the criminal act. But how would that fit within the linear inference model?

I have identified three complications which the linear inference model alone seems unable to address:

> THRESHOLD. How high should the source probability be? How strongly should the proposition 'source' be supported by the evidence?
>
> PRESUMPTION. Can we presume PRESENCE from SOURCE?
>
> LARGER CONTEXT. Why is PRESENCE an argument for GUILT?

I submit that, in order to address these difficulties, the linear model should be supplemented with what I call the *narrative-based model*. It is the topic of the next section.

8.4 THE BENEFITS OF CRIME NARRATIVES

I have suggested here and there that we cannot properly assess the probative value of DNA evidence unless we locate it in the larger context of trial proceedings. What is this larger context? In this section, I show that DNA evidence can be better assessed provided, first, the prosecutor offers an incriminating narrative, and second, the defense has an effective opportunity to scrutinize it. To make my argument more concrete, I shall use a fairly realistic, though still quite simplified, scenario:

> *Ennio*. A woman is found dead in the woods. The investigators recover remnants of semen on her body, which is severely wounded; they also recover blood stains in a parking lot near the woods. From the semen on the woman's body, a DNA profile is created. Forensic experts estimate that the DNA profile in question has a statistical frequency of 1 in 100 million. Through a database search, it turns out that an individual in the neighborhood, Ennio, has a matching DNA profile. Ennio is arrested and charged with murder.

8.4.1 The prosecutor's prima facie case

Imagine you are the prosecutor and you are constructing an incriminating case against Ennio. To a first rough approximation, four basic evidentiary facts can be identified:

(*semen*) Semen traces on the victim's body;

(*blood*) Blood stains in the parking lot;

(*wound*) Wounds from a knife on victim's body;

(*genetic match*) DNA match between Ennio and semen.

Again from the prosecutor's point of view, the four evidentiary facts above tend to prove four fact probanda:

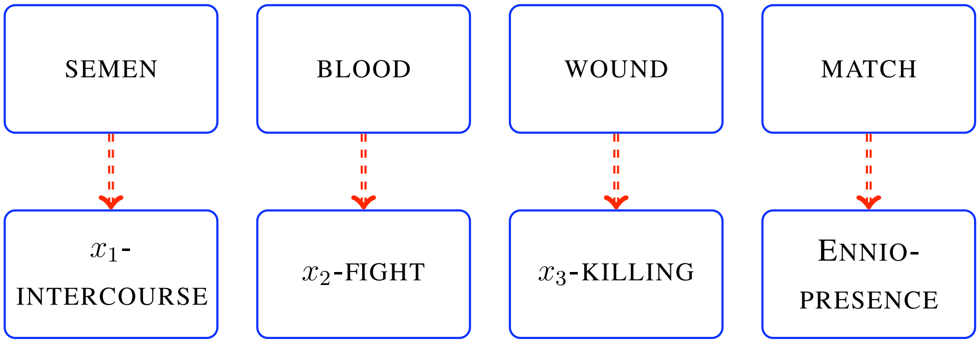
(x_1-*intercourse*) individual x_1 had intercourse with victim;

(x_2-*fight*) individual x_2 had violent fight with victim in parking lot;

(x_3-*killing*) individual x_3 killed victim with a knife;

(Ennio-*presence*) Ennio was present at the crime scene.

The prosecutor, in constructing his incriminating case, can advance the following four inferences:



The inference from the genetic match to Ennio's presence at the crime scene was discussed earlier, in section 8.3, by using the linear inference model. The new model I am articulating here builds on the linear model, so the reasoning used before can be applied again to warrant the prosecutor in advancing the provisional conclusion that Ennio was present at the crime scene. Recall that the linear model faced a a problem because it was not clear whether there was a presumption from source to presence. This problem can now be put aside because the conclusion that Ennio was present at the crime scene need not be established beyond a reasonable doubt; it is only put forward by the prosecutor as *prima facie* acceptable. To put it differently, we are now simply analyzing the matter within the prosecutor's standpoint. I am not yet concerned with whether the prosecutor can advance the proposition 'source' at trial.

The *prima facie* acceptability of the other provisional conclusions (i.e. intercourse; fight; killing) depends on a number of case-by-case details: whether the semen on the victim's body was in a quantity, location, and arrangement that indicate intercourse; whether the wound on the victim's body was caused by a human artefact, e.g. a knife, and whether it caused the victim's death; etc. Without dwelling on unnecessary details, let's assume that forensic experts, on the basis of their experience, are willing to claim that the conclusions in question can be *prima facie* drawn.

The next move for the prosecutor is to advance a *unification claim*:

$$x_1 = x_2 = x_3 = \mathit{Ennio}$$

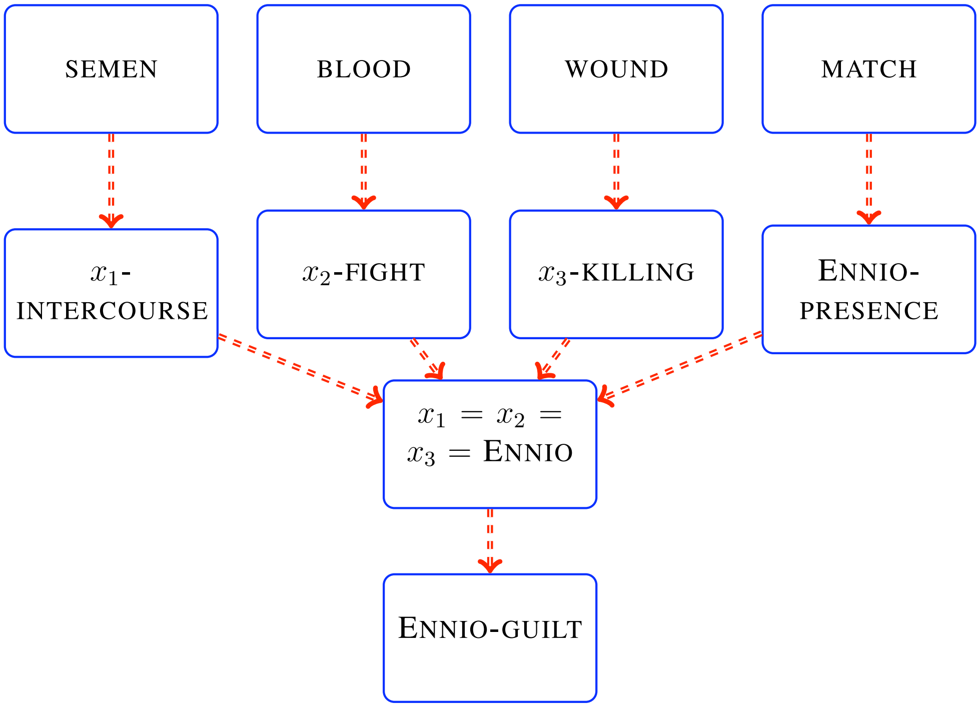
The four initial inferences proceeded in isolation. They still left open the possibility that whoever had intercourse with the victim might be a different person from whoever fought with or killed the victim. They still left open the possibility that different people participated in the crime. The unification claim, instead, rules out these possibilities and identifies Ennio as the (only? main?) perpetrator. The unification claim is very important because it entails that Ennio performed the *actus reus*. The claim allows the prosecutor to conclude that Ennio had sexual intercourse with the victim, fought with the victim and finally killed her with a knife.

With the unification claim in place, we begin to see a coherent and unifying narrative of what happens. What is the ground to assert the unification claim? Once again, I am not yet concerned with whether this claim has been established beyond a reasonable doubt. I am only concerned with whether the prosecutor can entertain such a claim with an initial degree of acceptability. A number of considerations can weigh in here: when examined by forensic experts, the crime traces suggest the presence of one perpetrator only; the crime could, in principle, be committed by one person only; Ennio would have had the physical force to commit the crime alone; etc. Depending on the information available here, the unification claim can turn out to be more or less strongly supported. (Note that the evidential support for the unification claim heavily rests on considerations of cohesiveness; see chapter 7 and the notion of *filling-in*.)

Putting everything together, the prosecutor's *prima facie* case against Ennio can be represented by the diagram on page 256. In plain words, the diagram can form the prosecutor's opening argument at trial, along the following lines:

> "The perpetrator had or attempted to have sexual intercourse with the victim in the parking lot (which explains the perpetrator's semen on the victim's body); a fight ensued during which the victim was wounded (which explains the blood stains in the parking lot); finally, the perpetrator killed the woman with a knife (which explains the wounds on the victim's body) and hid her body in the woods. Ennio is the perpetrator: he has a matching DNA profile whose frequency is as low as 1 in 100 millon."

The reader might wonder: How did we suddenly get to trial? Weren't we simply analyzing the prosecutor's internal mental processes in constructing a case against Ennio?



True enough. But we could get to trial because the prosecutor, at least in my reconstruction, assembled a case that, when it is considered in its entirety, advances a *prima facie* argument against Ennio.

8.4.2 Challenges and responses.

So far I have looked at the case from the prosecutor's standpoint. It is time to take up the defense's standpoint. The defense can challenge the prosecutor's reconstruction of the crime—i.e. what I'd like to call the *prosecutor's proposed narrative of the crime*—in many different ways; the possibilities to raise challenges are in principle infinite. Recall that the prosecutor, at least in my presentation of the case, relied on four crucial evidentiary facts: semen on the victim's body; blood in the parking lot; a wound on the victim's body; a genetic match linking Ennio to the crime scene. These four evidentiary facts licensed the *prima facie* acceptable conclusion that Ennio had intercourse with the victim, fought with her, and finally killed her.

(There is some ambiguity concerning the status of the challenges I am going to describe. They can be interpreted as challenges which the prosecutor himself foresees and against which he takes preemptive measures. They can also be interpreted as challenges which the defense actually raises at trial or could be thinking of raising at trial. In reading what follows, this ambiguity should be kept in mind.)

The defense could attack the prosecutor's case by arguing that Ennio had intercourse with the victim at a time that long preceded the killing; that the blood stains in the parking lot formed long before the victim's death; that the formation of a cut on the victim's body preceded the killing; that multiple people were involved in the crime; etc. Presumably, the defense will back these objections with some forensic findings, and to these objections the prosecutor will have to respond. If he failed to respond adequately, the defense would have exposed reasonable doubts in the incriminating case, such as: Did Ennio have intercourse and then killed the victim, or were the two events distant in time? How many people participated in the killing? Was Ennio alone? Could it be that Ennio had a fight with the victim, but others killed her? These questions, in turn, would raise other questions about the identity of the victim. Who was she? What was she doing in the parking lot? Why was Ennio there? For instance, if it turned out that the victim was a prostitute who was trying to break free from her pimp, it wouldn't be implausible that Ennio was only partly involved in the crime. Things would be different if Ennio and the victim were acquaintances, friends, lovers, partners, etc. The possibilities here are countless.

Each case is different, infinitely complex in details and variations. What should begin to appear is that many circumstances and details which we would have initially thought to be relatively unimportant—e.g. Ennio's habits and the victim's personal history—matter a lot for the solution of a case. This does not mean that legal fact-finding should turn into story-telling or gossiping about people's lives. Yet, besides some core evidential facts, such as forensic findings and DNA evidence, a myriad of secondary details and circumstances should not be ignored. The latter are as important as the former. This point makes good sense if we think of legal fact-finding as the prosecutor's attempt to offer a well-specified crime narrative which the defense can scrutinize and challenge. Let me illustrate this point by formulating a possible defense's objection which has more squarely to do with DNA evidence—an objection which will require considering seemingly secondary details and circumstances.

The defense could object that, even granted that Ennio genetically matched with the semen on the victim's body, he never visited the crime scene and he never had any encounter with the victim. We saw this objection already in a slightly different form, while discussing the linear model and the inferential step from *source* to *presence*. How could the defense sustain this objection? I can think of two ways to do so:

> (*fake genetic material*) First, the defense might argue that someone synthesized the genetic material and artificially implanted it in the crime scene.
>
> (*no link to victim or parking lot*) Second, the defense might argue that Ennio had no reason to be in contact with the victim, that there was no reason why Ennio would go to *that* parking lot.

If we were confined to the linear model, it would be unclear how to move forward. Do the objections mean that we cannot infer or presume *presence* from *source* at all? What the linear model tells us, at best, is that we can infer source if it is sufficiently probable given presence. But it is hard to say what 'sufficiently' is supposed to mean here. To avoid the impasse, we can view the two objections as what they are, namely objections to both of which the prosecutor is expected to respond. They are not frivolous objections, and the prosecutor should take them seriously. (On the notion of a frivolous objection, I will say more in the next parenthetic remark.)

In response to the defense's claim that fake genetic material was implanted, the prosecutor could perform the appropriate forensic testing. For there exists a specific forensic procedure that allows one to distinguish fake genetic material from real ones (Frumkin et al., 2009). If the forensic testing showed that the genetic material was not fake, this result would weaken the defense's position dramatically. So, the defense should think twice before raising this objection. If the defense is going to raise the fake-genetic-material objection, it must have good grounds to do so, e.g. it thinks that fake genetic material had, in fact, been implanted. (Suppose the defense had raised the objection that alien creatures implanted the fake genetic material and that no current human forensic technology could tell the difference between real and alien-fake genetic material. This would be a frivolous objection. The reason is that there would be no way for the prosecutor to respond and to adjudicate the matter. I am inclined to suggest, then, that a frivolous objection invites us to entertain a possible fact which is such that the adjudication of its truth is—in principle—beyond the reach of the prosecutor, the defense, the judge, or any party in the trial.)

Let's now turn to the second possible objection, i.e. that Ennio had no reason to visit the parking lot or be in contact with the victim. Anticipating this objection should prompt the prosecutor to seek more information about Ennio's habits and acquaintances. Who is Ennio? What kind of people does he know? Who are his acquaintances? What reasons would lead Ennio to go to the parking lot? In attempting to answer these questions, the prosecutor will most likely search for character evidence about Ennio. This, however, should not be a problem. Although character evidence is typically inadmissible to establish the commission of a crime, it should be admissible for other purposes.

The back and forth between prosecution and defense can continue, but the discussion above, I hope, should give a sense of how this dialectical process can unfold. What has emerged is that, besides a few core items of evidence, a myriad of initially secondary details can become quite important.

8.5 When is it enough, then?

Given the emphasis I've placed on well-specified crime narratives, the reader might have formed the impression that I have downplayed the importance of numbers and statistics in DNA cases. *I have not.* There is a significant difference between a DNA case in which the matching genetic profile has an estimated frequency of 1 in 50,000 and a DNA case in which the profile's estimated frequency is 1 in 1 billion. Or suppose that the case against Ennio were deprived of the DNA statistics. Then, suddenly, the incriminating case would lose its force almost entirely. There would be no evidence placing Ennio at the crime scene, and thus, the crime narrative constructed around this supposition would collapse.

A more general question now arises. When can we say that Ennio's guilt has been proven beyond a reasonable doubt? The answer I wish to suggest is based on two requirements:

> WELL-SPECIFIED NARRATIVE. The prosecutor should put forward an incriminating narrative, i.e. a narrative that, at least, answers the principal question (who did it? why? how? etc.) and that, in addition, answers many of the questions that naturally arise.
>
> RESISTANCE TO CHALLENGES. The prosecutor's narrative should resist all challenges coming from the defense lawyer who had an effective opportunity to scrutinize the prosecutor's narrative and who took full advantage of such an opportunity.

In chapter 7, I outlined what a well-specified narrative of the crime should look like. Let me recall the essentials here. A crime narrative is well-specified if it gives an account of the co-occurrence of the *mens res* and the *actus reus*. This means, for example, that the narrative should describe what motivated the accused to perform certain actions, what goals drove the accused, and what consequences resulted from those actions. The narrative that the prosecutor offered against Ennio has some of these features. It describes what motivated Ennio to kill, i.e. a fight following sexual intercourse, and it describes what consequence resulted, i.e. bloodstains in the parking lot and a deep, fatal cut on the victim's body. There might be lingering doubts regarding what motivated Ennio to do what he, presumably, did. Was it an incident or did he plan ahead? The defense could argue that the prosecutor's narrative does not give a sufficiently compelling account of the co-occurrence of *mens rea* and *actus reus*. Possibly so; the prosecutor might thus have to offer a more precise account of what happened. In this dialectic, a more compelling case against Ennio might result, or reasonable doubts might emerge which weaken the prosecutor's case. We cannot tell beforehand.

To stress once more the importance of a well-specified crime narrative, it is instructive to compare Ennio's case with a case in which a well-specified narrative is missing, despite the impressive DNA statistics:

> *Titus*. A woman is found dead in a California national park. The dead body was buried, but wild animals dug it out and fed on the flesh. On the woman's

> clothes, a few faded blood stains are visible, and from them, forensic experts create a DNA profile. They estimate that the DNA profile in question has a frequency of 1 in 50 billion. Through a database search, it turns out that an individual now living in Alaska, Titus, has a matching DNA profile. Titus is arrested.

Despite the impressive DNA statistics—i.e. the matching DNA profile has a frequency as low as 1 in 50 billion—the case against Titus is weaker than the case against Ennio. Titus lives in Alaska and presumably the murder happened in California. How did he get from one place to the other? And although a few blood stains were found on the victim's clothes, how does this finding have to do the victim's death? If such a case were to go to trial, the prosecutor would be expected to tell the fact-finders more. For instance, he should explain how the victim was killed and in what circumstances; how Titus was involved in the killing; when and how he flew from Alaska to California; etc. The impressive DNA statistics alone create a powerful case that Titus had some contact with the victim. But this is all they do. The DNA evidence simply "places" Titus in the surroundings of the victim's whereabouts for some reason or another. If the prosecutor's case were to rest on the DNA statistics and the DNA match only, it would be a rather weak case, far from establishing guilt beyond a reasonable doubt.

The fact is that, hopefully, a case like *Titus* won't go to court. Any sensible district attorney will use the DNA identification evidence to find further evidence, to discover why and how Titus got in touch with the victim, etc. An attorney will naturally look for more information and build a more cohesive incriminating case, one that resembles the case against Ennio. But let's suppose, for the sake of argument, that a case such as *Titus* would go to trial, and that a jury would have to decide it. If so, the presiding judge will be entitled to utter the following:

> I should, I think, members of the jury just sound a word of caution about the statistics. However compelling you may find them to be, we do not convict people in these courts on statistics. It would be a terrible day if that were so. Cited in (Kadane, 2008, p. 409)

If convicting people on mere statistics means anything at all, convicting Titus would be such an example. Ennio’s case, instead, is very different, for the DNA statistics are accompanied by a wealth of other considerations, items of information, and details. I have suggested in this chapter that the presence of a unifying narrative renders a conviction against Ennio more acceptable than a conviction against Titus.

8.6 ADDENDUM: COLD-HIT CASES

I now apply the narrative-based model to DNA cold-hit cases which have sparked some controversy in recent years. A suspect is typically identified by means of police investigations; this is the traditional, standard method. But more recently, identifications have also been carried out through database searches alone. Here is how it works: the investigators find some DNA traces at the crime scene, they create a DNA profile from the traces, and the resulting profile is then run through a database of existing profiles. If a DNA profile in the database shows a match, the individual associated with the matching profile becomes a suspect. Cases in which a suspect is identified through a database search are called cold-hit cases, and they are rising in number.

Cold-hit cases have triggered a controversy among legal scholars and forensic experts. Suppose the defendant is found to match and his DNA profile has an estimated frequency of 1 in 1 million; would it make any difference as to the significance of the match whether we are in a cold-hit case or in a standard case? There are two positions in the literature. The first asserts that in cold-hit cases a match has *less* significance; another asserts that it has the *same* significance. (Some even suggest that it has *more* significance.)

A parallelism can illustrate the claim that the match is less significant. In tossing a coin successively for ten times, a series of ten heads in a row is a quite rare and surprising event. It would be very surprising if, upon your very first attempt, you would get ten heads in a row. But now suppose you make 1,000 attempts, and on each attempt you toss the coin ten times successively. If among those 1,000 attempts you get at least one series of ten heads in a row, this result would not be so surprising any more. After all, getting ten heads in a row if you try enough times is practically certain. Now, what happens in a cold-hit case would seem to be very similar to what happens when you get ten heads in row after trying 1,000 times or more. A DNA profile is a very rare feature, so if you test only one individual and find a match, this is a very significant and surprising event. Yet, if you test 10 million people and find a match, the match becomes much less surprising. And this is what happens in cold-hit cases: you run the DNA profile against a database containing several million profiles, in the hope of finding a match. The analogy with the coin tossing case shows that, in cold-hit cases, the finding of a match is less surprising and thus less significant (NRC, 1996). To take this fact into account, the National Research Council suggests that a genetic profile's estimated frequency should be multiplied by the size of the database. This means that, for a database of size $1,000,000$ and a profile frequency of $1/10,000,000$, the adjusted frequency will be $1/10$; this is a much less significant figure than the initial $1/10,000,000$. Note that if the database is of size n and the genetic frequency is $1/n$, the adjusted frequency will be 1; this essentially means that the DNA evidence match has no significance at all.

Let's now consider the claim that a match in a standard or cold-hit case has always the same significance. Here are two arguments for this claim. *First*, the significance of a match is determined by the frequency of the DNA profile for which the match is declared. If the frequency is the same in the cold-hit scenario and in the standard scenario, the match has the same significance. Similarly, the chance of getting a series of ten heads in a row does not change depending on whether you make one attempt or millions of attempts. For each attempt, the chance of getting ten heads in a row is very low—i.e. $0.5^{10} = 0.0009765625$. *Second*, if one had a world database containing the DNA profile of each individual, the finding of a (unique) match would certainly be extremely significant, contrary to the NRC approach (Balding and Donnely, 1996).

(There are those who think that a match found in a cold-hit is more probative. The reason is that, since databases contain the profiles of previous offenders, the probability that a previous offender committed a crime is higher, hence a DNA match in a cold-hit case would be more probative than in standard cases (Macedo, 2008).)

What is the reason for the disagreement among forensic experts and statisticians? There seem to be two probability values here. One is the probability of getting ten heads in a row on any given attempt; this probability does not change and it equals $0.5^{10} = 0.0009765625$. Another is the probability that, in a sufficiently larger number of attempts, we get ten heads in a row at least once. This probability reaches 1 as the number of attempts increases. In DNA evidence cases, the two probabilities would be: the probability of finding a genetic match on any given attempt (which equals the profile's estimated frequency); the probability of finding (at least) one match in a large database. Balding and Donnely (1996) have in mind the former probability and the NRC has in mind the latter probability. Recall that the NRC suggested to adjust the profile's frequency by the size of the database. Now, in a database of size 1,000,000, the probability of finding (at least) one individual with a profile having a frequency of $1/1,000,000$ equals one, as the NRC adjustment predicts.[33]

As far as I can tell, in cold-hit cases, forensic experts are talking past each other. Some are interested in the probability of a match *on any given attempt to find a match*. Others are interested in the probability of a match *in a large database*. To be sure, forensic experts are genuinely disagreeing about which of the two probabilities is the more appropriate to assess the probative value of DNA evidence. The suggestion I want to make, building on the arguments of this chapter, is that we cannot appreciate the probative value of DNA evidence in isolation, as merely an issue concerning the significance of a genetic match. Let me explain.

In standard cases, we have at least two sources of incriminating evidence: the police investigation that led to the identification of the suspect; and the DNA match. In cold-hit cases, instead, we lack the additional information coming from the police investigation. One the one hand, this difference should not affect the estimated profile's frequency, so Balding and Donnely (1996) are partly right here. On the other, standard cases are different from cold-hit cases because in the former the police investigation was more thorough and extensive. The way to appreciate this difference is not, as the NRC does, by narrowly focusing on the significance of the genetic match in isolation from the rest of the incriminating evidence. Recall that the prosecutor, as illustrated in this chapter, has to provide a well-specified incriminating narrative, a narrative which can be supported by DNA evidence *as well as* other items of information. In standard cases, the information collected by the police before the database search can contribute to support the prosecutor's narrative. In cold-hit caes, since the information coming from the police is missing, the overall prosecutor's case will be weaker. I conclude that,

[33] Using the binomial distribution formula, the probability of finding *exactly one* matching individual is 0.625.

without focusing exclusively on the narrow question of a match's significance, the dispute among experts on cold-hit cases can disappear by having in sight the prosecutor's overall narrative and its role in the prosecutor's case.

CHAPTER 9

LOOKING BACK

It is time to recapitulate. In these concluding remarks, I reflect on the nature of statistical evidence and the role of probability in understanding the criminal standard of proof.

9.1 IS STATISTICAL EVIDENCE ENOUGH TO CONVICT?

My answer to this question has been evasive—and inevitably so. I think it is a misleading question to ask. To see why, let's look at DNA evidence, which is a type of evidence that rests crucially on statistics, i.e. the estimate of a DNA profile's frequency. In chapter 8, I argued that—*under some circumstances*—DNA evidence can be enough to sustain a criminal conviction. There I contrasted two scenarios, repeated here for convenience:

> *Ennio*. A woman is found dead in the woods. The investigators recover remnants of semen on her body, which is severely wounded; they also recover blood stains in a parking lot near the woods. From the semen on the woman's body, a DNA profile is created. Forensic experts estimate that the DNA profile in question has a statistical frequency of 1 in 100 million.

> Through a database search, it turns out that an individual in the neighborhood, Ennio, has a matching DNA profile. Ennio is arrested and charged with murder.
>
> *Titus*. A woman is found dead in a California national park. The dead body was buried, but wild animals dug it out and fed on the flesh. On the woman's clothes, a few faded blood stains are visible, and from them, forensic experts create a DNA profile. They estimate that the DNA profile in question has a frequency of 1 in 50 billion. Through a database search, it turns out that an individual now living in Alaska, Titus, has a matching DNA profile. Titus is arrested.

In both scenarios the DNA statistics link the suspect to the victim. Had the statistics been unavailable, the case against Ennio or Titus would not even have started. The difference, however, is that *Ennio* provides a well-specified narrative of what happened, whereas *Titus* does not. The narrative in *Ennio* is along the following lines:

> "The perpetrator had or attempted to have sexual intercourse with the victim in the parking lot (which explains the perpetrator's semen on the victim's body); a fight ensued during which the victim was wounded (which explains the blood stains in the parking lot); finally, the perpetrator killed the woman with a knife (which explains the wounds on the victim's body) and hid her body in the woods. Ennio is the perpetrator: he has a matching DNA profile whose frequency is as low as 1 in 100 millon."

Statistical evidence, whatever its form, should be used to construct a well-specified incriminating narrative. In giving evidential support to a narrative, statistical evidence will cohere with other items of information. So it would be difficult—perhaps even pointless—to demarcate sharply the contribution of the statistics to the narrative from the contribution of other items of information.[1] There are no criteria to individuate items of evidence. This is especially apparent in the case of DNA evidence. What is DNA evidence? Is it just a genetic match and a profile's frequency? What about the

[1] On this score, in chapter 8, sections 8.3 and 8.4, I have contrasted the *linear inference model* with a *narrative-based model*, and I have suggested that the latter affords us a better way to appreciate the probative value of DNA evidence.

shape and quantity of the genetic material? More generally, it is artificial to isolate items of evidence and attribute good or bad epistemic qualities to them, as though they existed independently *qua* things in themselves.

The difficulty in sharply demarcating pieces of evidence from others suggest that it is misleading to ask whether statistical evidence alone is enough to convict. The question implicitly assumes that if some types of statistical evidence are problematic, it must be because they have certain features which make them deficient. I reject this assumption. As suggested in chapter 1, I do not think there is anything wrong with statistics *per se*; what matters is the way in which they are used. Recall the contrast between *Ennio* and *Titus*. In the former, statistics were used appropriately, but in the latter, they were not. What marks the difference is that, in one case, a well-specified narrative was being offered, and in the other, the narrative was fragmentary, grossly incomplete, or even entirely missing.

But what is a narrative? In chapter 7, section 7.2, I sketched some ideas as to what a well-specified narrative should look like. A narrative is well-specified if it gives an account of the co-occurrence of the *mens rea* and the *actus reus*. So, for example, the narrative should describe what motivated the accused to perform certain actions, what goals drove the accused, and what consequences resulted from those actions; this is what I called an *event-narrative*. Another feature that makes a narrative well-specified is that it can accomodate (or "explain") the presence of certain crime traces, be those physical traces, cognitive traces left in the memory of a witness, or even digital traces left on a camera or recording device; this is what I called an *evidence-narrative*.

Now, any incriminating narrative will always have gaps, of one kind or another. To expect the prosecutor to specify everything would be too much. In chapters 7 and 8, I suggested that narrative specificity can be better understood within the adversarial process. In particular, taking into account the challenges which the defense might level against the prosecutor's narrative can help us recognize the appropriate degree of narrative specificity. Defense challenges are of two types: they can point out gaps in the narrative (*spoliation challenges*) or they can target specific assertions that the prosecutor's narrative makes about what happened (*challenges proper*). My account of when a narrative counts as reasonably well specified, then, depends on the lively adversarial confrontation of accusation and defense. In this process, a narrative is offered, challenged, refined, challenged again, and so on.

The importance of the adversarial process for discovering the appropriate degree of narrative specificity was illustrated at length in chapter 8 with respect to DNA evidence. As a second example, I shall briefly consider the *Shonubi* case from chapter 4. Here are the basic facts. Charles Shonubi was found carrying 427.4 grams of heroin at JFK airport in New York, having flown from Nigeria. He The travel records showed that Shonubi had made seven other trips between New York and Nigeria. On the basis of statistical evidence regarding other drug smugglers, it was very likely that Shonubi carried drugs on the other seven trips. On a conservative estimate, Shonubi carried at least 200 grams on each trip.

Presumably, the prosecutor's narrative in *Shonubi* was along the following lines: on certain dates Shonubi made 8 trips between the U.S. and Nigeria; on each trip he carried at least 200 grams of heroin; on the eighth trip he carried more than 400 grams. The statistical evidence, the travel records, and the airport search that occurred on Shonubi's eighth trip support the prosecutor's narrative with some degree of probability. Does the narrative in question count as sufficiently specific? It is not so clear how detailed a narrative should be in drug trafficking cases.[2] Should the narrative in these cases comprise details about, for example, who helped the suspect or whether he was part of a larger scheme, or is to enough to describe the quantity of drugs he carried and when?

I do not think we can reach any agreement on how specific the narrative against Shonubi should be by thinking abstractly about the case. Instead, we can make progress if we turn to the adversarial dimension of trial proceedings. On one hand, the prosecutor claims that Shonubi carried a certain amount of drugs, distributed over 8 smuggling

[2] Some might hold that it is enough to prove the defendant's possession of the illegal substance, while others might insist that the suspect's intentional participation in a planned scheme of drug trafficking should be proven as well. For instance, possibly because of the need to more effectively combat drug trafficking, in the state of Florida the *mens rea* is not even part of a drug offense. The *2002 Florida's Drug Abuse Prevention and Control Law*, 893.101(2), states that "knowledge of the illicit nature of a controlled substance is not an element of any offense under this chapter."

trips on a Nigeria-U.S. route. The defense could respond in different ways. If the prosecutor claims that *all* trips were drug smuggling trips, Shonubi could respond that *some* trips were family related. Alternatively, Shonubi might concede that all trips were drug smuggling trips, yet he could say that he carried very small quantities, below 200 hundred grams. Suppose the amount carried on each trip is the point of disagrement. If so, the prosecutor will be forced to make his proffered narrative more precise. What argument can the prosecutor offer that Shonubi carried 200 grams on *each* trip and not less? For all we know, the statistics suggest that smugglers rarely carry less than 200 grams. But why? One reason for this could be strictly economic: transporting small quantitates is not profitable. The prosecutor's narrative, then, could be enriched with the claim that Shonubi carried 200 grams on each trip because carrying less would not be profitable. Now, given this added detail, Shonubi could respond that carrying smaller quantities would still be profitable and that he did carry less on his seven trips because he was on training. The back-and-forth between prosecution and defense will continue until a fixed point is reached. In the end, depending on the details of the case, and on what arguments the prosecutor puts forward, Shounubi will have different defense strategies at his disposal.

The important point here is that the prosecutor need *not* say why exactly a certain amount was carried in each and every drug smuggling case. But on the supposition that Shonubi challenged the prosecutor's case by claiming that he carried less than 200 grams, the possibility that he carried less raises a reasonable doubt, and the prosecutor's narrative should address this doubt. Had Shonubi raised another challenge, the detail to be added to the narrative would be different. In this sense, the appropriate degree of narrative specificity cannot be fixed once and for all. Rather, it emerges through the adversarial process.

One might wonder at this point: are the statistics enough to conclude that Shonubi carried a total of roughly 2,000 grams of heroin although he was found carrying only 400 grams? I am inclined to think that *Shonubi* is not a clear-cut case and that much depends on the details. But my general point stands: there is nothing wrong with the statistics brought against Shonubi *per se*; what matters is the incriminating narrative that is being offered, where the appropriate degree of narrative specificity is discovered through the adversarial process. The enterprise of finding an inherent feature of statistical evidence which would make it deficient, irrespective of its contribution within a larger crime narrative, is hopeless. Statistical evidence, as any other evidence, has an inferential role (e.g. it supports a crime narrative); its probative value has to be assessed relative to its inferential role and on a case-by-case basis.

9.2 Is probability of guilt enough to convict?

I now turn to the question of whether the criminal standard of proof can be understood as a threshold guilt probability. The answer I defended throughout the dissertation, especially in chapters 2 and 7, is a resounding No. My argument has a negative part and a more constructive part. The negative part establishes that a probabilistic threshold need not be all there is to the criminal standard of proof. The constructive part identifies the additional features besides a probabilistic threshold.

The legal probabilists view the criminal standard of proof as an appropriately high threshold guilt probability. The reason for setting a high threshold in criminal cases is that the harm associated with convicting an innocent defendant is thought to outweigh the harm associated with acquitting a guilty defendant. As shown in chapter 2, a high probabilistic threshold has the function of *distributing errors* so as to minimize wrongful convictions at cost of increasing wrongful acquittals. What the legal probabilists do not consider is whether the criminal standard of proof can also have the function of *reducing errors*.

There are two reasons for disregarding error reduction and focusing on error distribution only. First, some trial errors will be inevitable no matter what. This is a consequence of the fallible nature of the trial system itself and of human evidence more generally. Second, pursuing a desirable error distribution (i.e. one that minimizes wrongful convictions at the cost of increasing wrongful acquittals) conflicts with error reduction. Setting the threshold for criminal convictions to a high probability value does promote a desirable error distribution, but only at cost of increasing the overall rate of error (see chapter 2, section 2.3). If so, the criminal standard of proof cannot promote both error reduction and a desirable error distribution; it can only promote one of the two. In selecting error distribution, the legal probabilists have assigned the function of promoting a desirable error distribution to an appropriately high threshold guilt probability.

I have argued, however, that we can isolate an aspect of error reduction that is not in direct conflict with pursuing a desirable error distribution. I called this aspect of error reduction the *trial system's ability to discriminate between innocent and guilty defendants*, or in short, the *discriminating power* of the trial system. Improving the trial system's discriminating power positively affects error reduction, but in ways that do not immediately clash with the equally important goal of pursuing a desirable error distribution. The goals of error reduction and of a desirable error distribution, then, can coexist in the criminal standard of proof without giving rise to incoherence. On one hand, an appropriately high threshold guilt probability promotes a desirable error distribution; on the other, maintaining the trial system's discriminating power above a certain level of performance promotes the equally important goal of error reduction. Error reduction is what is missing from the probabilist account of the criminal standard of proof.

I now turn to the more positive part of my argument, which was spelled out in the second part of chapter 2 and then developed more fully in chapter 7. Here is my proposed account of the criminal standard of proof:

> When a prosecutor attempts to establish guilt beyond a reasonable doubt, four requirements apply:
>
> NARRATIVITY. The prosecutor should put forward a complete (or quasi-complete) incriminating narrative, i.e. a narrative that, at least, answers the principal questions (who did it? why? how? etc.) and all (or most of) the event- and evidence-related questions which naturally arise.
>
> HIGH PROBABILITY. The prosecutor should put forward an incriminating narrative that is highly probable on the evidence.
>
> EVIDENTIAL COMPLETENESS. The prosecutor should make sure that, *relative to the proposed narrative*, the evidence presented at trial is complete, i.e. it is not affected by type one or type two gaps. If there are type one

> gaps in the evidence, the prosecutor should give a satisfactory explanation for them. If there are type two gaps in the evidence, the prosecutor can "fill in" the narrative (see the previous section for an account of filling-in).
>
> RESILIENCY. The prosecutor's narrative should resist all challenges coming from the defense lawyer who had an effective opportunity to scrutinize the prosecutor's narrative and who took full advantage of such an opportunity.

A plausible conjecture—which has been informing my discussion throughout chapters 2, 7 and 8—is that we can improve the discriminating power of criminal trials if the prosecution provides a well-specified narrative of the crime, if the evidence is as complete as possible, and if the defense levels challenges against the prosecutor's case. These are three components of my account of the criminal standard: narrativity; evidential completeness; and resiliency. If these three components serve the goal of error reduction by improving the system's discriminating power, the remaining component serves the function of distributing errors in a desirable way.

My four-pronged account of the criminal standard acknowledges the inevitability of errors, but it also underscores our ability to reduce errors by improving the overall discriminating power of the trial system. On my account, the criminal standard serves two functions: distributing errors by setting a threshold; demanding that the trial system's discriminating power be improved through narrativity, evidential completeness, and resiliency.

I should admit that I have not offerred a conclusive argument that narrativity, evidential completeness, and resiliency all positively contribute to the discriminating power of criminal trials. The reality of trial proceedings is not one of rational and disinterested cooperation in the search for the truth. Each party is more concerned with winning than with seeking the truth (Stuntz, 2011). As a prosecutor amasses evidence against the accused, he might do so for the sake of boosting his ego. As the defense introduces more evidence or raises challenges, it might introduce misleading evidence for the sake of winning, not for the sake of understanding what actually happened. The criminal justice system should be devised to maximize the incentives for producing non-misleading evidence and for formulating non-misleading arguments, and a probabilistic and game-theoretic analysis might be helpful here.[3] But all in all, there is no systemic cure against purposeful deception, bad faith, and human moral weaknesses, unless we appeal to the moral integrity of the participants. The trial participants have the moral duty to do their best in seeking the truth and nothing but the truth. If they purposely choose to do otherwise and deceive their opponent, it is tempting to say, as our *ultima spes* and *refugium peccatorum*, that they should forever be damned in the flames of hell.

[3] For a game-theoretic defense of the right to silence, see (Seidmann and Stein, 2000).

APPENDIX A

A.1 THE MCCLESKEY DILEMMA

Drawing inferences from statistical data is difficult and controversial. The existence of purposeful discrimination is often established by means of statistica data, especially in cases involving employment and wage discrimination. Discrimination, however, is not a central issue in criminal cases and thus it received almost no mention in this dissertation. In this appendix, I want to spend a few words on the issue. I do so by comparing two cases. They show an ambivalent and almost contradictory attitude toward statistical evidence by the U.S. Suprem Court.

McCleskey, a black man, was convicted of murder on October 12, 1987. The jury recommended that McCleskey be sentenced to death. On appeal, the Supreme Court of Georgia affirmed the conviction and the death penalty recommendation. McCleskey filed a petition for habeas corpus in the Federal District Court of the Northern District of Georgia. He complained that the Georgia capital sentencing process discriminated against blacks in violation of the 8th and the 14th Amendement. McCleskey relied on a very careful and sophisticated statistical study, known as the *Baldus study* (Baldus et al., 1990). The study examined over 2,000 murder cases that occurred in Georgia during the 1970's. Here are the raw numbers:

> (-) defendants charged with killing white persons received the death penalty in 11 percent of the cases;

> (-) defendants charged with killing blacks received the death penalty in only 1 percent of the cases; and
>
> (-) the death penalty was administered in 22 percent of the cases involving black defendants and white victims; 8 percent of the cases involving both white defendants and white victims; and 3 percent of the cases involving black defendants and black victims.

The Baldus study took into account 230 variables that could have explained the data in a non-discriminatory way. Baldus concluded that defendants charged with killing white victims were 4.3 times more likely to receive the death penalty than defendants charged with killing blacks. On this statistical grounds, McCleskey argued that the Georgia capital sentencing process was openly discriminatory against blacks.

The District Court, the 11th Circuit, and the U.S. Supreme Court were unpersuaded. They found nothing wrong with the Georgia capital sentencing process, or to put it more carefully, they were not persuaded that the Baldus study showed any purposeful discrimination against McCleskey. The Supreme Court wrote:

> Each jury is unique in its composition, and the Constitution requires that its decision rest on consideration of innumerable factors that vary according to the characteristics of the individual defendant and the facts of the particular capital offense. McCleskey v. Kemp, 481 U.S. 279 at 294 (1987).
>
> [...] At most, the Baldus study indicates a discrepancy that appears to correlate with race. Apparent disparities in sentencing are an inevitable part of our criminal justice system. *McCleskey* at 299.

The Court held that statistical evidence alone could at best establish a widespread correlation between race and death penalty recommendations. The reason here is that the statistics could not establish a purposeful discrimination against McCleskey because jurors' decisions are inevitably discretionary. According to the Court, establishing the existence of purposeful discrimination against McCleskey would require the consideration of more variables and factors than the relatively few which the Baldus study considered. In other words, the Court thought that *the Baldus study lacked individualization* (on the notion of individualized evidence, see chapters 1).

Justice Brennan disagreeed. In his dissent, he wrote:

> . . . a death sentence must be struck down when the circumstances under which it has been imposed "creat[e] an unacceptable risk that 'the death penalty [may have been] meted out arbitrarily or capriciously' or through 'whim or mistake' " (quoting California v. Ramos, 463 U.S. 992, 999 (1983)). *McCleskey* at 323.

The majority thought that what needed to be established was the purposeful discrimination specifically against McCleskey. The majority concluded that the Baldus study did not establish that. In contrast, Brennan thought that what needed to be established was the statistical, generalized, or systemic risk that the Georgia sentencing process discriminated against blacks. Brennan thought, and I believe the majority would agree, that the Baldus study established the existence of such a systemic risk. But which is to be established? The individual instance of discrimination, or the systemic risk of discrimination? This is a particularly difficult question which I should be content only to raise.

What is puzzling about the McCleskey decision is that one year earlier the U.S. Supreme Court reached an apparently contradictory conclusion on the subject of statistical evidence as proof of discrimination.

Employees of the North Carolina Agricultural Extension Service filed suit against various state and local officials for racial discrimination in employment, under title VII of the Civil Right Act of 1964. At trial, petitioners relied on regression analysis to show that blacks were paid less than similarly situated white employees. The regression showed that in 1974 black employees earned $331 less per year than similarly situated white employees. Against respondent's argument that petitioners did not take into account all the relevant variables, the Court wrote:

> While the omission of variables from a regression analysis may render the analysis less probative than it otherwise might be . . . it is clear that a regression analysis that includes less than "all measurable variables" may serve to prove a plaintiff's case. A plaintiff in a Title VII suit need not prove discrimination with scientific certainty; rather, his or her burden is to prove discrimination by a preponderance of the evidence. Bazemore v. Friday, 478 U.S. 385 at NN (1986).

Compared to the McCleskey decision, it looks as though the Court here reached a diametrically opposite conclusion. There the Court reasoned that more variables needed to be taken into account, while here the Court reasoned that it was not necessary to take into account more variables. It is worth emphasising that the Baldus study took many more variables into the consideration than the much simpler regression analysis in Bazemore. If so, why did the Court reached a different, and seemingly contradictory, conclusion? In *McCleskey*, while referring to *Bazemore*, the Court wrote:

> Thus, the application of an inference drawn from the general statistics to a specific decision in a trial and sentencing simply is not comparable to the application of an inference drawn from general statistics to a . . . Title VII case. In those cases, the statistics relate to fewer entities, and fewer variables are relevant to the challenged decisions. *McCleskey* at 295.

The Court's reasoning is that, in a Title VII employment discrimination case, the relevant variables are fewer than in a sentencing discrimination case. But this raises once more the dilemma I mentioned earlier. In a death penalty discrimination case, what should be established? Is it the specific jury discrimination or the systemic risk of jury discrimination? I wanted to compare *McCleskey* and *Bazemore* to show that the U.S. Supreme Court itself is somewhat ambiguous in its attitudes toward statistical evidence. This is a difficult issue, and in my view, there is no easy way to reconcile the two cases. In employment discrimination cases, it is clear that what is to be established is the systemic discrimination, not the employee-specific, discrimination. But why should that not be so in death penalty discrimination cases, as well?

BIBLIOGRAPHY

Peter Achinstein. Concepts of evidence. *Mind*, 87(345):22–45, 1978.

Ronald J. Allen. A reconceptualization of civil trials. *Boston University Law Review*, 66:401–437, 1986.

Ronald J. Allen. On the signficance of batting averages and strikeout totals: a clarification on the“naked statistical evidence” debate, the meaning of “evidence” and the requirement of proof beyond a resonable doubt. *Tulane Law Review*, 65:1093–1100, 1991.

Ronald J. Allen. How presumptions should be allocated: Burdens of proof, uncertainity, and amibiguity in modern legal discourse. *Harvard Journal of Law and Public Policy*, 17:627–646, 1994.

Ronald J. Allen. No plausible alternative to a plausible story of guilt as the rule of decision in criminal cases. 2010.

Ronald J. Allen and Brian Leiter. Naturalized epistemology and the law of evidence. *Virginia Law Review*, 87:1491–1550, 2001.

Ronald J. Allen and Michael S. Pardo. The myth of the law-fact distinction. *Northewestern University Law Review*, 97(4):1769–1808, 2003.

Ronald J. Allen and Michael S. Pardo. The problematic value of mathematical models of evidence. *Journal of Legal Studies*, 36(1):107–140, 2007.

Terence Anderson, David A. Schum, and William Twining. *Analysis of Evidence (2nd Edition)*. Cambridge University Press, 2005.

David J. Balding. *Weight-of-evidence for forensic DNA profiles*. John Wiley and Sons, 2005.

David J. Balding and Peter Donnely. Evaluating DNA profile evidence when the suspect is identified through a database search. *Journal of Forensic Science*, 41:603–607, 1996.

David Baldus, Charles Pulaski, and George Woodworth. *Equal Justice and the Death Penalty: A Legal and Empirical Analysis*. Northwestern University Press, 1990.

V. C. Ball. The moment of truth: Probability theory and stanfords of proof. *Vanderbilt Law Review*, 14:807–830, 1961.

Charles L. Barzun. Rules of weight. *Notre Dame Law Review*, 83(5):1957–2018, 2007.

Cesare Beccaria. *Dei delitti e delle pene*. 1764.

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

Jacob Bernoulli. *Ars conjectandi*. 1713.

Luc Bovens and Stephan Hartmann. Solving the riddle of coherence. *Mind*, 112:601–633, 2003.

John Bradshaw and G. Wallance. Models for the processing and identification of faces. *Perception and Psychophysics*, 9(5):443–448, 1971.

John Buckleton. A framework for interpreting evidence. In John Buckleton, Christopher M. Triggs, and Simon J. Walsh, editors, *Forensic DNA Evidence Interpretation*. CRC Press, 2005a.

John Buckleton. Population genetic models. In John Buckleton, Christopher M. Triggs, and Simon J. Walsh, editors, *Forensic DNA Evidence Interpretation*. CRC Press, 2005b.

Robert P. Burns. The distinctiviness of trial narratives. In Antony Duff, Lindsay Farmer, Sandra Marshall, and Victor Tadros, editors, *The Trial on Trial (VOL 1): Truth and Due Process*. Hart Publishing, 2004.

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

Gianfranco Carofiglio. *Ragionveoli dubbi*. Sellerio, 2006.

Ronald H. Clark, William S. Bailey, and George R. Dekle. *Cross Examination Handbook: Persuasion Strategies and Techniques*. Aspen Publishers, 2010.

Jonathan L. Cohen. *The Probable and the Provable*. Oxford University Press, 1977.

Jonathan L. Cohen. Subjective probability and the paradox of the gatecrasher. *Arizona State Law Journal*, pages 627–634, 1981a.

Jonathan L. Cohen. Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences*, 3:317–370, 1981b.

Stewart Cohen. Knowldge and context. *Journal of Philosophy*, 83(10):574–583, 1986.

Mark Colyvan, Helen M. Regan, and Scott Ferson. Is it a crime to belong to a reference class? *Journal of Political Philosophy*, 9(2):168–181, 2001.

Alan D. Cullison. Probability analysis of judicial fact-finding: A preliminary outline of the subjective approach. *Toledo Law Review*, 1:538–598, 1969.

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

G.M. Davies, H. D. Ellis, and J. W. Shepherd. Face recognition accuracy as a function of mode presentation. *Journal of Applied Psychology*, 63:180–187, 1978.

Philip Dawid. Bayes's theorem and weighing evidence by juries. In *Bayes's Theorem*, volume 113, pages 71–90. Oxford University Press, 2002.

Bruno de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937.

Michael L. Dekay. The difference between blackstone-like error ratios and probabilistic standards of proof. *Law and Social Inquiry*, 21:95–132, 1996.

Keith DeRose. Knowledge, assertion and lotteries. *Australia Journal of Philosophy*, 76: 568–580, 1996.

Keithe DeRose. Assertion, knowledge and context. *The Philosohical Review*, 111(2): 167–203, 2002.

Alan M. Dershowitz. *Reasonable Doubts: The Criminal Justice System and the O.J. Simposon Case*. Simon and Schuster, 1997.

Dennis J. Devine. *Jury Decision Making: The State of the Science*. New York University Press, 2012.

Keith Devlin. Scientfic heat about cold hit (manuscript). 2007.

Persi Diaconis and Frederick Mosteller. Methods for studying coincidences. *Journal of the American Statistical Association*, 84:853–861, 1989.

Igor Douven. A new solution to the paradoxes of rational acceptance. *British Journal of Philosophy of Science*, 53:391–410, 2002.

Igor Douven and Timothy Williamson. Generalizing the lottery paradox. *British Journal of Philosophy of Science*, 57:755–779, 2006.

Fred Dretske. Conclusive reasons. *Australian Journal of Philosophy*, 49:1–22, 1971.

Itiel E. Dror, David Charlton, and Ailsa E. Peron. Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, 156:74–78, 2006.

Daniel Ellesberg. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75(4):643–669, 1961.

David Enoch, Talia Fisher, and Levi Spectre. Statistical evidence, sensitivity, and the legal value of knowledge. *Philosophy and Public Affairs*, 40(3):197–224, 2012.

Lina Erikson and Alan Hájek. What are degrees of belief? *Studia Logica*, 86:183–213, 2007.

Stephen E. Feinberg and David H. Kaye. Legal and statistical aspects of some mysterious clusters. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154:61–74, 1991.

Stephen E. Fienberg and Miron L. Straf. Statistical evidence in US courts: An appraisal. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 145:49–59, 1991.

Michael O. Finkelstein and William B. Fairley. A Bayesian approach to identification evidence. *Harvard Law Review*, 83(3):489–517, 1970.

Michael O. Finkelstein and William B. Fairley. A comment on “trial by mathematics”. *Harvard Law Review*, 84:1801–1809, 1971.

Michael O. Finkelstein and Bruce Levin. *Statistics for Lawyers (second edition)*. Springer, 2001.

George Fisher. *Evidence (2nd Edition)*. Foundation Press, 2008.

Branden Fitelson. Likelihoodism, Bayesianism, and relational confirmation. *Synthese*, 16(47):1–22, 2006.

George F. Fletcher. *Basic Concepts of Crimnal Law*. Oxofrd University Press, 1998.

James Franklin. *The Science of Conjecture: Evidence and Probability before Pascal*. John Hopkins University Press, 2001.

David A. Freedman. What is the chance of an earthquake? Technical Report 611, UC, Berkeley, Department of Statistics, 2003.

Richard D. Friedman. Route analysys of credibility and hearsay. *The Yale Law Journal*, 97(4):667–742, 1987.

Richard D. Friedman. Standards of persuation and the distinction betwen fact and law. *Northewestern University Law Review*, 86:916–942, 1992.

Richard D. Friedman. A presumption of innocence, not of even odds. *Stanford Law Review*, 52:873–887, 2000.

Dan Frumkin, Adam Waserstrom, Ariane Davidson, and Arnon Grafit. Authentication of forensic dna samples. *Forensic Science International: Genetics*, 4(2):95–103, 2009.

Maria C. Galavotti. *Philosophical Introduction to Probability*. Stanford University Press, 2005.

Maria C. Galavotti. Probability, statistics, and law. In D. Dieks, W. Gonzalez, S. Hartmann, M. Stoeltzner, and M. Weber, editors, *Probability, Laws, and Structures*, pages 402–412. Springer, 2012.

Peter Gärdenförs, Bengt Hansson, Nils-Eric Sahlin, and Sören Halldén. *Evidentiary Value: Philosophical, Judicial, and Psychological Aspects of a Theory. Essays Dedicated to Sören Halldén on his Sixtieth Birthday*. C.W.K. Gleerup, 1983.

Donald G. Gifford. The challenge to the individual causation requirement in mass products torts. *Washington and Lee Law Review*, 63:873–935, 2005.

Gerd Gigerenzer, Peter M. Todd, and the ABC Research Group. *Simple Heuristics That Make Us Smart*. Oxford University Press, 1999.

Alvin Goldman. Discrimination and perceptual knowledge. *Journal of Philosophy*, 73: 771–791, 1976.

Jean Goodwin. Wigmore's chart method. *Informal Logic*, 20(3):223–243, 2000.

Lisa Kern Griffin. Narrative, truth, trial. *Georgetown Law Journal*, 101:281–335, 2013.

Susan Haack. *Evidence and Inquiry*. Wiley-Blackwell, 1995.

Susan Haack. Legal probabilism: An epistemological dissent (manuscript). 2011.

Ian Hacking. *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge University Press, 1984.

Ian Hacking. *An Introduction to Probability and Inductive Logic*. Cambridge University Press, 2001.

Alan Hájek. The reference class problem is your problem too. *Synthese*, 156(3):563–585, 2007.

Alan Hájek. Interpretations of probability. *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*, 2010.

Gilbert Harman. Knowledge, inference, explanation. *American Philosophical Quartely*, 5:164–173, 1968.

Leon Hart and Tony Honore. *Causation in the Law (second edition)*. Oxford University Press, 1985.

John Hawthorne. *Knowledg and Lotteries*. Oxford University Press, 2004.

Martin H. Hewett. A more reliable right to present a defense: The compulsory process clause after Crawford v. Washington. *Georgetown Law Journal*, 96:273–315, 2007.

Hock Lai Ho. *Philosophy of Evidence Law*. Oxford University Press, 2008.

Rinke Hoestra and Joost Breuker. Commonsense causal explanation in the legal domain. *Artificial Intelligence and Law*, 15:281–199, 2007.

Joachim Hruschka. *Die Konstitution des Rechtsfallen. Studien zum Verhaltniss von Tatsachenfeststellung und Rechtsanwendung*. Duncker and Humblot, 1965.

Louis L. Jaffe. Res ipsa loquitur vindicated. *Insurance Law Journal*, 1(6):130–139, 1952.

Geroge F. James. Relevancy, probability and the law. *California Law Review*, 29(6): 689–705, 1941.

Stuart H. James, Paul E. Kish, and T. Paulette Sutton. *Principles of Bloodstains Pattern Analysis: Theory and Practice*. CRC Press, 2005.

R. N. Jonakin. When blood is their argument: Probabilities in criminal cases, genetic markers, and once again, Bayes' theorem. *University of Illinois Law Review*, pages 369–421, 1983.

Joseph B. Kadane, editor. *Statistics in the Law*, 2008. Oxford University Press.

Daniel Kahneman and Amos Tversky. Evidential impact of base rates. pages 153–160, 1982.

Daniel Kahnemman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, XLVII(263-291), 1979.

John Kaplan. Decision theory and the fact-finding process. *Stanford Law Review*, 20 (1065-1092), 1968.

John Kaplan, Robert Weisberg, and Guyora Binder. *Criminal Law: Cases and Materials (6th edition)*. Aspen Publishers, 2008.

David H. Kaye. Probability theory meets res ipsa loquitur. *Michican Law Review*, 77: 1456–1484, 1978.

David H. Kaye. The paradox of the gatecrasher and other stories. *The Arizona State Law Journal*, pages 101–110, 1979a.

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

David H. Kaye. Review: Naked statistical evidence. *The Yale Law Journal*, 89(3): 601–611, 1980.

David H. Kaye. Paradoxes, gedanken experiments and the burden of proof: A response to dr. cohen's reply. *Arizona State Law Journal*, pages 635–646, 1981.

David H. Kaye. Limits of the preponderance of the evidence standard: Justifiably naked statistical evidence and multiple causation. *The American Bar Foundation Research Journal*, pages 487–516, 1982.

David H. Kaye. The admissibility of probability evidence in criminal trials – part I. *Jurimetrics Journal*, 26:343–346, 1986a.

David H. Kaye. Do we need a calculus of weight to understand proof beyond a reasonable doubt? *Boston University Law Review*, 66:657–672, 1986b.

David H. Kaye. The admissibility of probability evidence in criminal trials – part II. *Jurimetrics Journal*, 27:160–172, 1986c.

David H. Kaye. The probability of an ultimate issue: The strange cases of paternity testing. *Iowa Law Review*, 75:75–109, 1989.

David H. Kaye. DNA evidence: Probability, population genetics and the courts. *Harvard Journal of Law and Techonology*, 7:101–172, 1993.

David H. Kaye. Clarifying the burden of persuasion: What Bayesian rules do and not do. *International Commentary on Evidence*, 3:1–28, 1999.

David H. Kaye. The non-science of fingerprinting: United States v. Llera-Plaza. *Quinnipiac Law Review*, 21:1073, 2003.

David H. Kaye. Probability, individualization, and uniqueness in forensic science evidence: Listening to the academics. *Brooklyn Law Review*, 75(4):1174–1186, 2010a.

David H. Kaye. *The Double Helix and the Law of Evidence*. Harvard University Press, 2010b.

David H. Kaye. Beyond uniqueness: the birthday paradox, source attribution and individualization in forensic science. *Law, Probability and Risk*, 12(1):3–11, 2013.

David H. Kaye and George F. Sensabaugh. Reference guide on DNA evidence. In *Reference Manual on Scientific Evidence (2dn ed.)*, pages 576–585. Federal Judicial Center, 2000.

Mark Kelman. *The Heuristics Debate*. Oxford University Press, 2011.

Jonathan J. Koehler. Error and exageration in the presentation of DNA evidence in trial. *Jurimetrics Journal*, 34:21–39, 1993.

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

Jonathan J. Koehler. The base rate fallacy reconsidered: descriptive, normative, and methodological challanges. *Behavioral and Brain Sciences*, 19(1):1–53, 1996b.

Jonathan J. Koehler. When do courts think base rate statistics are relevant? *Jurimetrics Journal*, 42:373–402, 2002.

Jonathan J. Koehler and Michael J. Saks. Individualization claims in forensic science: Still unwarranted. *Brooklyn Law Review*, 75(4):1187–1208, 2010.

Jonathan J. Koehler and Daniel N. Shaviro. Veridical verdicts: Increasing verdict accuracy through the use of overtly probabilistic evidence and methods. *Cornell Law Review*, 75:247–279, 1990.

Jonathan J. Koehler, A. Chia, and J. S. Lindsey. The random match probability (RMP) in DNA evidence: Irrelevant and prejudicial? *Jurimetrics Journal*, 35:201–219, 1995.

Igal Kvart. A probabilistic theory of knowledge. *Philosophy and Phenomenological Research*, 72:1–43, 2006.

Pierre-Simon Laplace. *Essai philosophique sur les probabilités*. 1814.

Larry Laudan. *Truth, Error, and Criminal Law: An Essay in Legal Epistemology*. Cambridge University Press, 2006.

Larry Laudan. The elementary epistemic arithmatic of criminal justice. *Journal of Social Epistemology*, 5:282–294, 2008.

Krista Lawlor. *Assurance: An Austinian view of Knowledge and Knowledge Claims*. Oxord University Press, 2013.

Hannes Leitgeb. Reducing belief simpliciter to degrees of belief (manuscript). 2010.

Richard O. Lempert. Modeling relevance. *Michigan Law Review*, 75(5/6):1021–1057, 1977.

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

Richard O. Lempert. The suspect population and dna identification. *Jurimetrics Journal*, 34:1–7, 1993.

David Lewis. Causation. *Journal of Philosophy*, 70:556–567, 1973.

David Lewis. Elusive knowledge. *Australia Journal of Philosophy*, 74:549–567, 1996.

Richard C. Lewontin. Comment: The use of DNA profiles in forensic contexts. *Statistical Science*, 9:259–262, 1994.

Charles E. Lindblom. The science of muddling through. *Public Administration Review*, 19(2):79–88, 1959.

Andrew Lingertwood. Can DNA evidence alone convict an accused? *Sydney Law Review*, 33:487–514, 2011.

Matthew R. Lippman. *Contemporary Criminal Law: Concepts, Cases, and Controversis (2nd edition)*. Sage Publication, 2010.

Elizabeth F. Loftus. *Eyewitness Testimony (revised edition)*. Harvard University Press, 1996.

Carmen De Macedo. Guilt by statistical association: Revisiting the prosecutor's fallacy and the interrogator's fallacy. *Journal of Philosophy*, 105(5):320–332, 2008.

Ronald Meester, Marieke Collins, Richard Gill, and Michiel van Lambalgen. On the (ab)use of statistics in the legal case against the nurse Lucia de B. *Law, Probability and Risk*, 2007.

Miguel A. Méndez. *Evidence: The California Code and the Federal Rules (4th edition)*. Thomson West, 2008.

Julia Mortera and Philip Dawid. Probability and evidence. In T. Rudas, editor, *Handbook of Probability Theory*. Sage Handbook, 2007.

Dana N. Nelkin. The lottery paradox, knowledge, and rationality. *The Philosophical Review*, 109:373–409, 2000.

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

Charles R. Nesson. The evidence of the event? on judicial proof and the acceptability of verdicts. *Harvard Law Review*, 98(1357-1392), 1985.

Robert Nozick. *Philosophical Explanations*. Harvard University Press, 1981.

NRC. *DNA Technology in Forensic Science*. National Academy Press, 1992.

NRC. *The Evaluation of Forensic DNA Evidence*. National Academy Press, 1996.

NRC. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press, 2009.

Michael S. Pardo and Ronald J. Allen. Judicial proof and the best explanation. *Law and Philosophy*, 27(223-268), 2008.

Nancy Pennington and Reid Hastie. A cognitive theory of juror decision making: the story model. *Cardozo Law Review*, 13:519–557, 1991.

Federico Picinali. Structuring inferential reasoning in criminal fact finding: An analogical theory. *Law, Probability and Risk*, 11(2/3):197–223, 2012.

Simeon Denis Poisson. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile.* 1837.

Karl Popper. *Logik der Forshung.* Springer, 1935.

Karl Popper. *The Logic of Scientific Discovery [English translation].* Routledge, 2002.

Richard Posner. *The Economic Analysis of Law.* Brown & Company, 1973.

Duncan Pritchard. *Epistemic Luck.* Oxford University Press, 2005.

Amit Pundik. *Statistical Evidence: In search of a Principle.* PhD thesis, University of Oxford, Faculty of Law, 2009.

Amit Pundik. The epistemoloigy of statistical evidence. *The International Journal of Evidence and Proof*, 15:117–143, 2011.

Frank P. Ramsey. Truth and probability. In R.B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter 7, pages 156–198. Kegan, Paul, Trench, Trubner and Co. Ltd, 1931.

Mike Redmayne. The relevance of bad character. *Cambridge Law Journal*, 61(3):684–714, 2002.

Mike Redmayne. Exploring the proof paradoxes. *Legal Theory*, 14:281–309, 2008.

Hans Reichenbach. *The Theory of Probability (English edition)*. University of Caifornia Press, 1949.

Bernard Robertson and G. A. Vignaux. Probability - the logic of the law. *Oxford Journal of Legal Studies*, 13:457–478, 1993.

Bernard Robertson and G. A. Vignaux. DNA evidence: Wrong answers or wrong questions? *Genetica*, 96:145–152, 1995.

Andrea Roth. Safety in numbers? Deciding when DNA alone is enouhg to convict. *New York University Law Review*, 85:1130–1185, 2010.

Sherrilyn Roush. *Tracking Truth: Knowledge, Evidence, and Science*. Oxford University Press, 2006.

Richard M. Royall. *Statistical Evidence: A likelihood paradigm.* Chapman and Hall/CRC, 1997.

Bertrand Russell. *Human Knowledge*. Simon and Schuster, 1948.

Michael J. Saks and Robert F. Kidd. Human information processing and adjudication: Trial by heuristics. *Law and Society Review*, 15(123-160), 1980.

Michael J. Saks and Jonathan J. Koehler. The individualization fallacy in forensic science evidence. *Vanderbilt Law Review*, 61:199–219, 2008.

Wesley Salmon. *Scientific Explanation and the Causal Structure of the World.* Princeton University Pressenton, 1984.

Chris W. Sanchirico. Character evidence and the object of trial. *Columbia Law Review*, 101(6):1227–1311, 2001.

Chris W. Sanchirico. What makes the engine go? Cognitive limitation and cross-examination. *Widener Law Review*, 14:507–524, 2009.

Boaz Sangero and Mordechai Halpert. Why a conviction should not be based on a single piece of evidence: A proposal for reform. *Jurimetrics Journal*, 48, 2007.

Frederick Schauer. *Profiles, Probabilities, and Stereotypes*. Belknap Press, 2003.

Richard Schmalbeck. The trouble with statisticla evidence. *Law and Contemporary Problems*, 49(3):221–236, 1986.

Leila Schneps and Coralie Colmez. *Math on Trial: How Numbers Get Used and Abused in the Courtroom*. Basic Books, 2013.

Ferdinand Schoeman. Statistical vs. direct evidence. *Nous*, 21(2):179–198, 1987.

David A. Schum and Joseph D. Kadane. *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. Wiley-Interscience, 1996.

David S. Schwartz. A foundation theory of evidence. *Georgetown Law Journal*, 100-171:95, 2011.

Daniel J. Seidmann and Alex Stein. The right to silence helps the innocent: A game-theoretic analysis of the fifth amendment privilege. *Harvard Law Review*, 114:430–510, 2000.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

Barbara Shapiro. *Beyond Reasonable Doubt and Probable Cause: Historical Perspectives on the Anglo-American Law of Evidence*. University of Caifornia Press, 1991.

Barbara J. Shapiro. *A Culture of Fact: England, 1500-1720*. Cornell University Press, 2003.

Daniel N. Shaviro. Statistical-probability evidence and the appearence of justice. *Harvard Law Review*, 103:530–554, 1989.

Dan Simon. A third view of the black box: Cognitive coherence in legal decision making. *The University of Chicago Law Review*, 71:511–586, 2004.

Rita James Simon and Lidha Mahan. Quantifying burdens of proof: A view from the bench, the jury, and the classroom. *Law and Society Review*, 5(3):319–330, 1971.

Daniel J. Simons and Christopher F. Chabris. Gorillas in our minds: Sustained inattention blidness for dynamic events. *Perception*, 28:1059–1074, 1999.

Kenneth W. Simons. Statistical knowledge deconstructed (manuscript). 2011.

Brian Skyrms. Resiliency, propensity, and causal necessity. *Journal of Philosophy*, 74: 704–713, 1977.

Brian Skyrms. *Causal Necessity*. Yale University Press, 1980.

Brian Skyrms. *Choice and Chance: An Introduction to Inductive Logic (4th edition)*. Wadsworth Publishing, 1999.

Steven Sloman. *Causal Models: How People Think About the World and Its Alternatives*. Oxford University Press, 2005.

John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., 2000.

Alex Stein. *Foundations of Evidence Law*. Oxford University Press, 2005.

Davod A. Stoney. Measurement of fingerprint individuality. In Henry C. Lee and Robert E. Gaensslen, editors, *Advances in Fingerprint Techonology (2nd edition)*. CRC Press, 2001.

William J. Stuntz. *The Collapse of The American Criminal Justice System*. Cambridge University Press, 2011.

Richard Swinburne, editor. *Bayes's Theorem*. Oxford University Press, 2002.

James W. Tanaka and Martha J. Farah. Parts and whole in face recognition. *The Quarterly Journal of Experimental Psychology*, 46A(3):225–245, 1993.

Franco Taroni and Colin G.G. Aitken. Forensic science at trial. *Jurimetrics Journal*, 37: 327–337, 1997.

Franco Taroni, Colin G.G. Aitken, Paolo Garbolino, and Alex Biederman. *Bayesian Networks and Probabilistic Inference in Forensic Science*. John Wiley and Sons, 2006.

Michele Taruffo. *La semplice verità*. Laterza, 2009.

Paul Thagard. *Cohrence in Thought and Action*. MIT Press, 2000.

Sandra Guerra Thompson. Beyond a reasonable doubt? reconsidering uncorroborated eyewitness identification testimony. *UC Davis Law Review*, 41:1487–1545, 2008.

William C. Thompson and Simon A. Cole. Psychological aspects of forensic identification evidence. In M. Costanzo, D. Krauss, and K. Pezdek, editors, *Expert Psychological Testimony in the Courts*, pages 31–68. Erlbaum, 2007.

William C. Thompson and Edward L. Shumann. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behaviour*, 11:167–187, 1987.

William C. Thompson, Franco Taroni, and Colin G.G. Aitken. How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Science*, 48: 47–54, 2003.

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

Peter Thomson. Margaret thatcher: A new illusion. *Perception*, 9(4):483–484, 1980.

Peter Tillers. United States v. Shonubi: A statistical oddity? *Cardozo Law Review*, 1997.

Peter Tillers and Jonathan Gottfried. A collateral attack on the legal maxim that proof beyond a reasonable doubt is unquantifiable. *Law, Probability and Risk*, 5:135–157, 2007.

Peter Tillers and Eric D. Green, editors. *Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism*, 1988. Springer.

Petr Tillers. Trial by mathematics – reconsidered. 2011.

Laurence H. Tribe. A further critique of mathematical proof. *Harvard Law Review*, 84 (8):1810–1820, 1971a.

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

Amos Tversky and Daniel Kahnemman. Judgment under uncertainity: Heuristics and biases. *Science*, 185:1124–1131, 1974.

Amos Tversky and Daniel Kahnemman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 40:293–315, 1983.

William Twining. *Theories of Evidence: Bentham and Wigmore*. Stanford University Press, 1986.

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

Jonathan Vogel. The new relevant alternatives theory. *Nous*, 33:155–180, 1999.

Richard von Mises. *Probability, Statistics and Truth (revised English edition)*. Macmillan, 1957.

Douglas N. Walton. *Legal Argumentation and Evidence*. Penn State University Press, 2002.

David Wasserman. Forensic DNA typing. In Justine Burley and John Harris, editors, *Companion to Genetics*. Blackwell, 2008.

David T. Wasserman. The morality of statistical proof and the risk of mistaken liability. *Cardozo Law Review*, 13:935–976, 1991.

Brian Weatherson. From classical to intuitionistic probability. *Notre Dame Journal of Formal Logic*, 44:111–123, 2003.

Bruce S. Weir. The rarity of DNA profiles. *The Annals of Applied Statistics*, 1:358–370, 2007.

Robert Weisberg. Proclaiming trials as narratives: Premises and pretenses. In Peter Brooks and Paul Gewirtz, editors, *Law's Stories: Narrative and Rhetoric in the Law*, pages 61–83. Yale University Press, 1996.

Francis L. Wellman. *The Art of Cross Examination*. The Macmillan Company, 1903.

Gary L. Wells. Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology*, 62:793–752, 1992.

Gary L. Wells, Amina Memon, and Steven D. Penrod. Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, 7(2):45–75, 2006.

James Q. Whitman. *The Origins of Reasonable Doubt: Theological Roots of the Criminal Trial*. Yale University Press, 2008.

Thomas D. Wickens. *Elementary signal detection theory*. Oxford University Press, 2002.

John Henry Wigmore. *The Principles of Judicial Proof (as Given by Logic, Psychology, and General Experience and Illustrated in Judicial Trials)*. Little, Brown, and Company, 1913.

Timothy Williamson. *Knowledge and Its Limits*. Oxford University Press, 2000.

James Woodward. *Making Things Happen: a Theory of Causal Explanation*. Oxford University Press, 2003.

Richard W. Wright. Causation, responsability, risk, probability, naked statistics, and proof: Pruning the bamble bush by clarifying the concepts. *Iowa Law Review*, 73: 1001–10077, 1988.

Richard W. Wright. Liability for possible wrongs: Causation, statistical probability, and burden of proof. *Loyola of Los Angeles Law Review*, 41:1295–1343, 2008.

Sandy L. Zabell. Fingerprint evidence. *Journal of Law and Policy*, 13:143–179, 2005.

Adrian A. S. Zuckerman. Law, fact or justice. *Boston University Law Review*, 66: 487–508, 1986.