Burdens of Proof - Sample Chapter

Marcello Di Bello and Rafal Urbaniak

1/26/2021

Introduction

After the evidence has been presented, examined and cross-examined at trial, trained judges or lay jurors must reach a decision. The decision criterion is defined by law and consists of a standard of proof, also called the burden of persuasion. So long as the evidence against the defendant is sufficiently strong to meet the requisite proof standard, the defendant should be found liable. This chapter begins with a brief description of standards of proof in the law, then outlines different probabilistic accounts of standards of proof, and finally discusses some objections to this approach.

Legal background

In criminal proceedings, the governing standard is 'proof beyond a reasonable doubt.' If the decision makers are persuaded beyond a reasonable doubt that the defendant is guilty, they should convict, or else they should acquit. In civil cases, the standard is typically 'preponderance of the evidence'. The latter is less demanding than the former, so the same body of evidence may be enough to meet the preponderance standard, but not enough to meet the beyond a reasonable doubt standard. A vivid example of this difference is the 1995 trial of O.J. Simpson who was charged with murdering his wife. He was acquitted of the criminal charges, but when the family of the victim brought a lawsuit against him, they prevailed. O.J. Simpson did not kill his wife according to the beyond a reasonable doubt standard, but he did according to the preponderance standard. An intermediate standard, called 'clear and convincing evidence', is sometimes used for civil proceedings in which the decision is particularly weighty, for example, a decision whether someone should be committed to a hospital facility.

This tripartite distinction of proof standards—beyond a reasonable doubt; preponderance; clear and convincing evidence—is common in Anglo-american jurisprudence. It is not universal, however. Different countries may use different standards. France, for example, uses the standard of 'intimate conviction' for both civil and criminal proceedings. Judges deciding cases 'must search their conscience in good faith and silently and thoughtfully ask themselves what impression the evidence given against the accused and the defence's arguments have made upon them' (French Code of Criminal Procedure, art. 353). German law is similar. Germany's Code of Civil Procedure, Sec. 286, states that 'it is for the court to decide, based on its personal conviction, whether a factual claim is indeed true or not.'

How to define standards of proof, or whether they should be even defined in the first place, remains contentious [@diamond90, @newman1993, @Horowitz1996,@laudan2006truth, @walen2015]. Judicial opinions offer different paraphrases, sometimes conflicting, of what these standards mean. The meaning of 'proof beyond a reasonable doubt' is the most controversial. It has been equated to 'moral certainty' or 'abiding conviction' (Commonwealth v. Webster, 59 Mass. 295, 320, 1850) or to 'proof of such a convincing character that a reasonable person would not hesitate to rely and act upon it in the most important of his own affairs' (US Federal Jury Practice and Instructions, 12.10, at 354, 4th ed. 1987). But courts have also cautioned that there is no need to define the term because 'jurors know what is reasonable and are quite familiar with the meaning of doubt' and attempts to define it only 'muddy the water' (U.S. v. Glass, 846 F.2d 386, 1988).

Probability theory can bring conceptual clarity to an otherwise heterogeneous legal doctrine, or at least this is the position of legal probabilists.

Probability-based rules of decision

Legal probabilists have proposed to interpret proof beyond a reasonable doubt as the requirement that the defendant's probability of guilt, given the evidence presented at trial, meet a threshold, say, >95%. Variations of this view are common [see @Bernoulli1713Ars-conjectandi, @Laplace1814, @kaplan1968decision, @Dekay1996, @kaye79, @laudan2006truth]. This interpretation is, in some respects, plausible. From a legal standpoint, the requirement that guilt be established with high probability, still short of 100%, accords with the principle that proof beyond a reasonable doubt is the most stringent standard of all but at the same time 'does not involve proof to an absolute certainty' and thus 'it is not proof beyond any doubt' (R v Lifchus, 1997, 3 SCR 320, 335). That this intepretation is quite natural is further attested by the fact that the probabilistic interpretation is taken for granted in psychological studies about people's understanding of proof beyond a reasonable doubt [@dhamiEtAl2015]. This research examines whether people use a 75% or 95% threshold, and does not question whether the standard functions as a probabilistic threshold.

Reliance on probabilistic ideas is even more explicit in the standard 'preponderance of the evidence'—also called 'balance of probabilities'—which governs decisions in civil disputes. This standard can be interpreted as the requirement that the plaintiff—the party making the complaint against the defendant—establish its version of the facts with greater than 50% probability. The 50% threshold, as opposed to a more stringent threshold of 95% for criminal cases, reflects the fact that preponderance is less demanding than proof beyond a reasonable doubt. The intermediate standard 'clear and convincing evidence' is more stringent than the preponderance standard but not as stringent as the beyond a reasonable doubt standard. Since it lies in between the other two, it can be interpreted as the requirement that the plaintiff establish its versions of the facts with, say, 75-80% probability.

Some worry that a mechanical application of numerical thresholds would undermine the humanizing function of trial decision-making. As [@tribe71] put it, '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, [their] humanizing function.' Thresholds, however, can vary depending on the costs and benefits at stake in each case (see later discussion). So they need not be applied mechanically without considering the individual circumstances (CITE Hedden and Colyvan, 2019). Furthermore, if jurors are numerically literate, they should not lose sight of their humanizing function as they would no longer be intimated by numbers. So the force of the objection underscores the need to ensure that jurors are numerically literate, not to dispense with numerical thresholds altogether.

Even if numerical thresholds cannot be used in the daily business of trial proceedings, they can still serve as theoretical concepts to understand the role of proof standards in the justice system, such as regulating the relative frequency of false positive and false negative decisions or minimizing expected costs. A more stringent threshold will decrease the number of false positives (say false convictions) at the cost of increasing the number of false negatives (say false acquittals), and a less stringent threshold will increase the number of false positives while decreasing the number of false negatives. This trade-off has been described, among others, by Justice Harlan in his concurring opinion In re Winship, 397 U.S. 358, 397 (1970). Against this background, it is natural to ask what the optimal or most efficient threshold should be. The optimal threshold may be one that minimizes false positives and false negatives overall or one that minimizes expected costs. Which threshold would minimize overall errors? Which would minimize expected costs? As shown below, these questions can be answered using the formal apparatus of probability theory, in combination with calculus and expected utility theory.

Minimizing expected costs

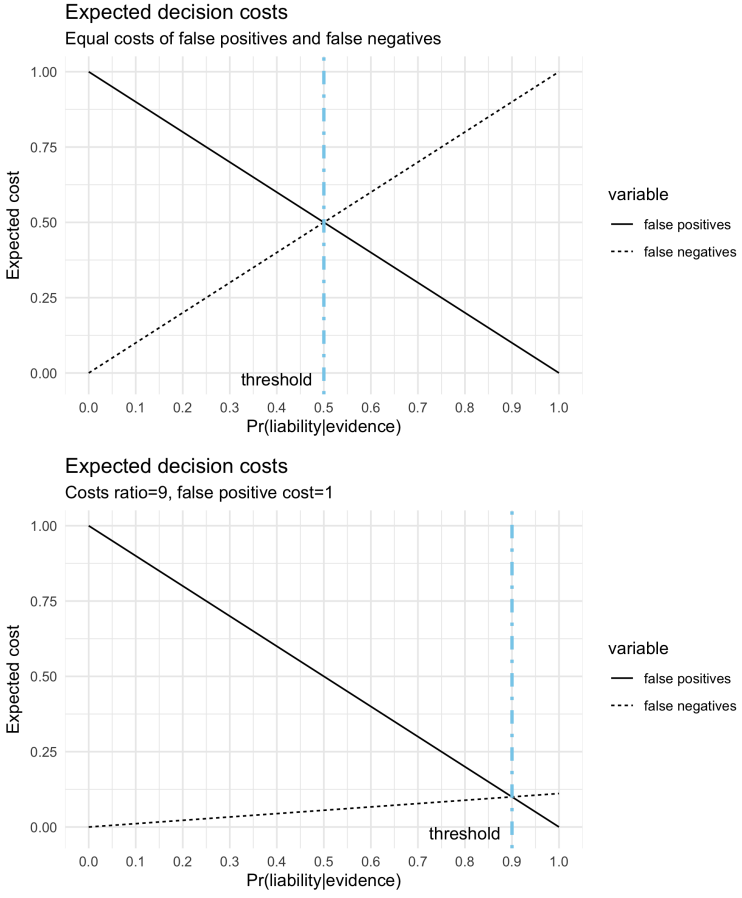
Probabilistic stadard of proof can be conceptualized through the lenses of expected utility theory [@kaplan1968decision, @Dekay1996, @hamer2004}. Broadly speaking, expected utility theory recommends agents to take the course of action that, among the available alternatives, maximizes expected utility. On this view,

the standard of proof is met whenever the expected utility (or cost) of a decision against the defendant (say, a conviction) is greater (or lower) than the expected utility (or cost) of a decision in favor of the defendant (say, an acquittal). Let $c(CI)$ be the cost of convicting a factually innocent defendant and $c(AG)$ the cost of acquitting a factually guilty defendant. For a conviction to be justified, the expected cost of convicting an innocent—that is, $c(CI)$ discounted by the probability of innocence $[1 - \Pr(G|E)]$—must be lower than the expected cost of acquitting a guilty defendant—that is, $c(AG)$ discounted by the probability of guilt $\Pr(G|E)$. This holds just in case

$$\frac{\Pr(G|E)}{1 - \Pr(G|E)} > \frac{c(CI)}{c(AG)}.$$

This inequality captures how high the probability of guilt or civil liability must be to justify a verdict against the defendant. If the cost ratio is 9—as might be appropriate in a criminal case—the inequality holds only if $\Pr(G|E)$ meets a 90% threshold.

The same analysis *mutatis mutandis* applies to civil cases in which mistaken decisions comprise mistaken attributions of liability (false positives) and mistaken failures to attribute liability (false negatives). If the cost ratio is one—as might be appropriate in a civil case in which false positives and false negatives are equally harmful—the inequality holds only if the probability that the defendant is civilly liable meets a 50% threshold.



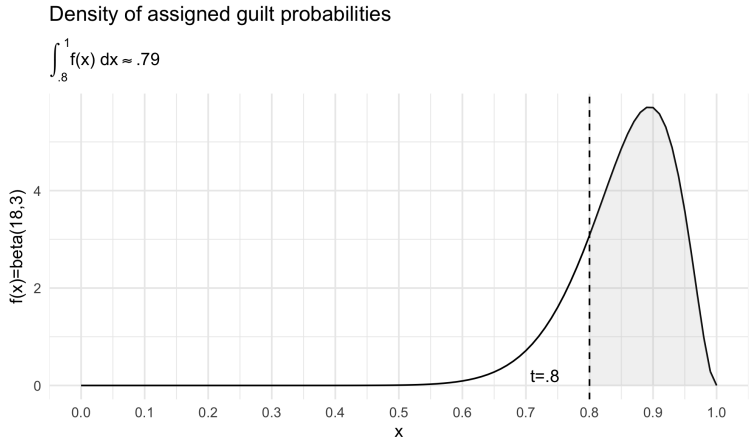
This analysis only considers the costs of mistaken decisions, but leaves out the benefits associated with correct decisions. More comprehensive analyses would consider both. The basic insight remains the same, however. The probability required for a conviction or a finding of civil liability against the defendant is a function of weighing the costs and benefits that would result from true and false positive as well as true and false negative decisions. On this account of proof standards, the stringency of the threshold depends on costs and

benefits, and thus different cases may require different thresholds. Cases in which the charge is more serious than others—say, murder compared to petty theft—may require higher thresholds so long as the cost of a mistaken decision against the defendant is more significant. Standards of proof would vary depending on the costs at stake in different cases. Whether or not standards of proof should vary in this way is a matter of debate [@kaplow2012, @picinali2013}. The same standard of proof is typically applied for murder and petty theft. The law typically makes coarse distinctions between standards of proof, such as 'proof beyond a reasonable doubt' for criminal cases, 'preponderance of the evidence' for civil cases and 'clear and convincing evidence' for a narrow subset of civil cases in which the accusation against the defendant is particularly serious. Another complication is that eliciting costs and benefits that result from trial decisions is not easy. Should they be elicited through a democratic process or should different jurors or judges apply their own in a subjective fashion? (CITE WHAT?) No matter the answer to these questions, when probabilistic standards of proof are paired with expected utility theory, they become part of the calculus of utilities. In line with the law and economics movement, trial decision-making is viewed as one instrument among others for maximizing overall social welfare [@Posner1973].

Minimizing overall errors

Instead of thinking in terms of maximizing expected utility (or minimizing expected costs), probabiListic standardS of proof can be also be viewd more directly as regulating the rate of erroneous trial decisions. We will see, however, that the error-centered approch agrees to a large extent with the approach based on maximizing expected utility.

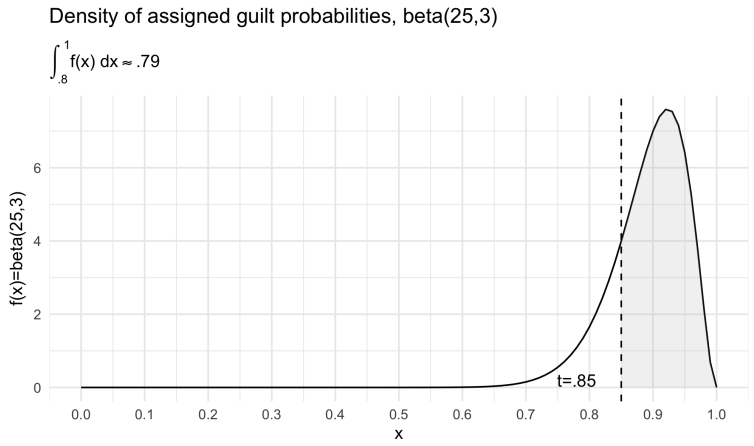
Consider an idealized model of a criminal trial system.
Each defendant is assigned a probability x of criminal liability (or guilt) based on the evidence presented at trial. Since over a period of time many defendants face charges, the guilt probability will have its own distribution. Extreme guilt probabilities set at 0% or 100%, presumably, are assigned rarely in trials if ever, while values between 40% and 80% are more common.
A rigorous way to express this distribution is by means of a probability density function, call it $f(x)$.
The figure below uses a right skewed distribution beta(18,3).



The right skew reflects the assumption that defendants in criminal cases are sent to trial only if the incriminating evidence against them is strong. It should be no surprise that most defendants are assigned a high probability of guilt. The distribution of the probability of liability in civil cases over a period of time might look quite different, probably centered around 50% or 60%.

In the figure above, the threshold for conviction is set at $> 80\%$, and the area under the curve to the right of the threshold is about .79. In other words, according to this model, 79% of defendants on trial are convicted and 21% acquitted. These figures are close to the rates of conviction and acquittal in many countries. Since $f(x)$ is a probability density, the total area under the curve adds up to 100%, encompassing all defendants, both convicted and acquitted defendants.

If the threshold becomes more stringent—for example, it moves up to 85%—the rate of conviction would decrease provided the underlying distribution does not change. But if the distribution becomes more skewed toward the right—say beta(25,3)—the rate of conviction could still be about 79% even with a more stringent threshold of 85%.

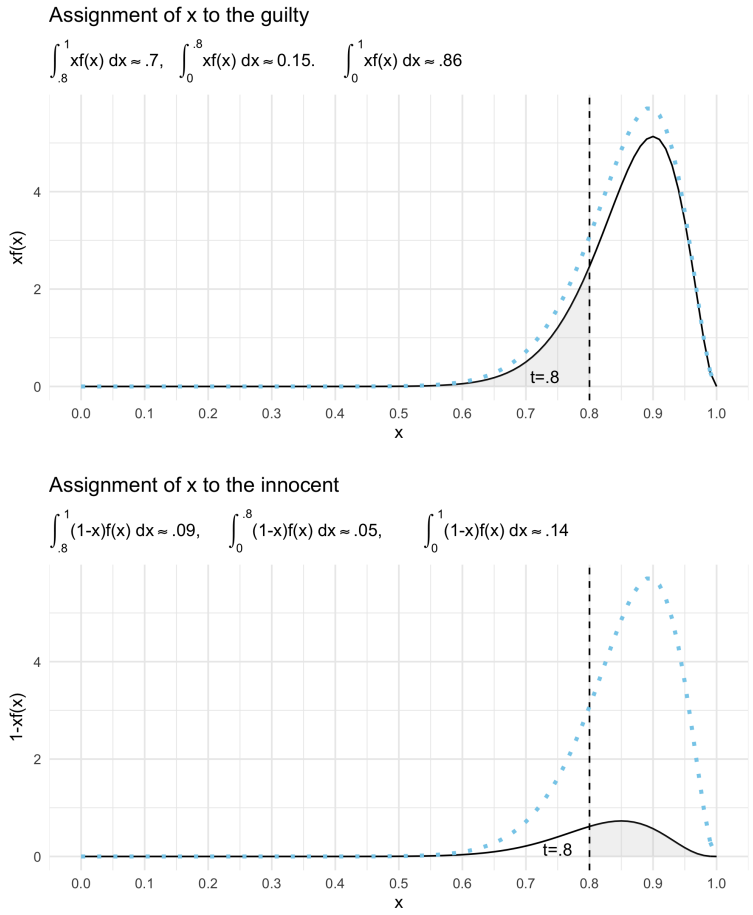


So far the model only describes the rate at which defendants are convicted depending on the strigency of the threshold. But it is also possible to represent in the model the rate at which innocent and guilty defedants are convicted.

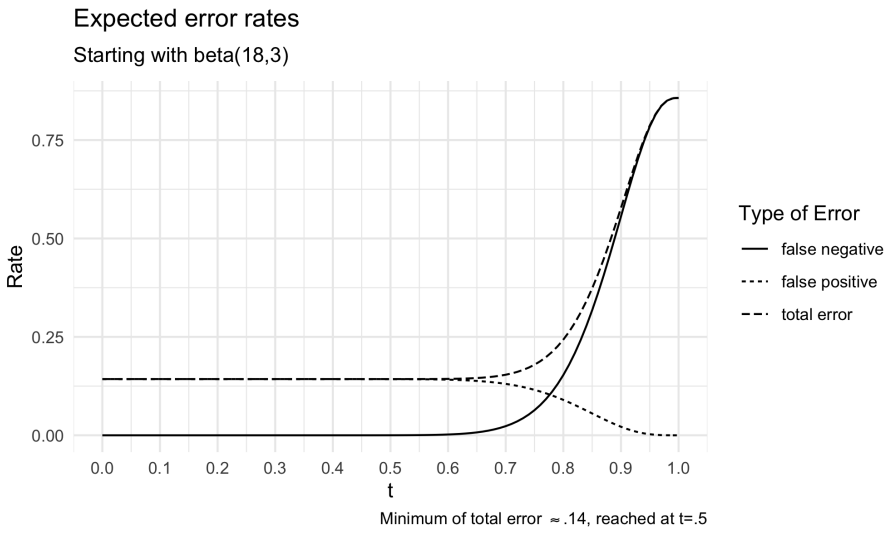
Presumably, among the defendants facing trail, some are factually innocent and the rest are factually guilty. What is the proportion of innocent and guilty defendants among all defendants? The expected proportion of guilty and innocent defendants on trial, out of all defendants, can be inferred from the density distribution $f(x)$ under certain assumptions. Suppose each defendant is assigned a guilt probability based on the best and most complete evidence. From the perspective of judges and jurors (or anyone who has access to the evidence and evaluates it the same way), x% of defendants who are assigned x% guilt probability are expected to be guilty and $(1-x)$% innocent. For example, 85% of defendants who are assigned a 85% guilt probability are expected to be guilty and 15% innocent; 90% of defendants who are assigned a 90% guilt probability are expected to be guilty and 10% innocent; and so on.

Consequently, the function $xf(x)$ describes the (expected) assignment of guilt probabilities for guilty defendants, and similarly, $(1-x)f(x)$ the (expected) assignment of guilt probabilities for innocent defendants. Neither of these functions is a probability density, since $\int_0^1 xf(x)\,\mathrm{d}x = 0.86$ and $\int_0^1 (1-x)f(x)\,\mathrm{d}x = 0.14$. That is, the total areas under the curve are, respectively, .86 and .14 (see graphs below). These numbers express the (expected) proportion of guilty and innocent defendants out of all defendants on trial, respectively 86% and 14%.

The rates of incorrect decisions—false convictions and false acquittals or more generally false positives and false negatives—can be inferred from this model as a function of the threshold t [@hamer2004, @hamer2014]. The integral $\int_0^t xf(x)\,\mathrm{d}x$ equals the expected rate of false acquittals, or in other words, the expected proportion of guilty defendants who fall below threshold t (out of all defendants), and the integral $\int_t^1 (1-x)f(x)\,\mathrm{d}x$ equals the expected rate of false convictions, or in other words, the expected proportion of innocent defendants who fall above threshold t (out of all defendants). The rates of correct decisions—true convictions and true acquittals or more generally true positives and true negatives—can be inferred in a similar manner. The integral $\int_t^1 xf(x)\,\mathrm{d}x$ equals the expected rate of true convictions and $\int_0^t (1-x)f(x)\,\mathrm{d}x$ the expected rate of true acquittals. In the figure below, the regions shaded in gray correspond to false negatives (false acquittals) and false positives (false convictions). The remaining white regions within the solid black curve correspond to true positives (true convictions) and true negatives (true acquittals). Note that the dotted blue curve is the original overall distribution for all defendants.



The size of the grey regions in the figures above—which correspond to false positives and false negatives—is affected by the location of threshold t. As t moves upwards, the rate of false positives decreases but the rate of false negatives increases. Conversely, as t moves downwards, the rate of false positives increases but the rate of false negatives decreases. This trade-off is inescapable so long as the underlying distribution is fixed. Below are both error rates—false positives and false negatives—and their sum plotted against a choice of t, while holding fixed the ensity function binom(18,3). The graph shows that any threshold that is no greater than 50% would minimize the total error rate (comprising false positives and false negatives). A more stringent threshold, say $> 90\%$, would instead significantly reduce the rate of false positives but also significantly increase the rate of false negatives, es expected.



In general, the threshold that minimizes the expected rate of incorrect decisions overall, no matter the underlying distribution, lies at 50%. The claim that setting threshold at $t = .5$ minimizes the expected error rate for any underlying distribution of x is general and holds for $t = .5$ only. It holds given the distribution $f(x)$ =beta(18,3) as well as any other distribution [@kaye1982limits, @Kaye1999Clarifying-the-, @cheng2015]. To show this, let $E(t)$ as a function of threshold t be the sum of rates of false positive and false negative decisions:

$$E(t) = \int_0^t xf(x)\,\mathrm{d}x + \int_t^1 (1-x)f(x)\,\mathrm{d}x.$$

The overall rate of error is minimized when $E(t)$ is the lowest. To determine the value of t for which $E(t)$ is the lowest, set the derivative of $E(t)$ %and $R(t)$ to zero, that is, $\frac{d}{dt}E(t) = 0$. By calculus, $t = 1/2$.[1] So a threshold of 50% is the one that minimizes the aggregate rate of erroneous decisions.

This claims holds when the two decisional errors are assigned the same weight, or in other words, the costs of false positives and false negatives are symmetric. The $> 50\%$ threshold therefore should be most suitable for civil trials. In criminal trials, however, false convictions are typically considered significantly more costly than false acquittals, say a cost ratio of 9:1 [but see @epps2015]. The sum of the two error rates can be weighted by their respective costs:

$$E(t) = \int_0^t xf(x)\,\mathrm{d}x + 9\int_t^1 (1-x)f(x)\,\mathrm{d}x.$$

Given a cost ratio of 9:1, the optimal threshold that minimizes the (weighted) overall rate of error is no

[1]Note that $\frac{d}{dt}E(t)$ is the the sum of the derivatives of $\int_0^t xf(x)\,\mathrm{d}x$ and $\int_t^1 (1-x)f(x)\,\mathrm{d}x$, that is,

$$\frac{d}{dt}E(t) = \frac{d}{dt}\int_0^t xf(x)\,\mathrm{d}x + \frac{d}{dt}\int_t^1 (1-x)f(x)\,\mathrm{d}x.$$

By the fundamental theorem of calculus,

$$\frac{d}{dt}\int_0^t xf(x)\,\mathrm{d}x = tf(t) \text{ and } \frac{d}{dt}\int_t^1 (1-x)f(x)\,\mathrm{d}x = -(1-t)f(t).$$

By plugging in the values,

$$\frac{d}{dt}E(t) = tf(t) - (1-t)f(t).$$

Since $\frac{d}{dt}E(t) = 0$, then $tf(t) = (1-t)f(t)$ and thus $t = 1-t$, so $t = 1/2$ or a $> 50\%$ threshold.

longer 1/2, but rather, $t = 9/10 = 90\%$.[2] Whenever the decision threshold is more stringent than $> 50\%$, the overall (unweighted) error minimization may be sacrificed to pursue other goals, for example, protecting more innocents against mistaken convictions, even at the cost of making a larger number of mistaken trial decisions overall.

The standard 'proof beyond a reasonable doubt' is often paired with the Blackstone ratio, the principle that it is better that ten guilty defendants go free rather than even just one innocent be convicted. The exact ratio is a matter of controversy [@voloch1997]. It is tempting to think that, say, a 99% threshold guarantees a 1:99 ratio between false convictions and false acquittals. But this would be hasty for at least two reasons. First, probabilistic thresholds affect the expected rate of mistaken decisions. The actual rate may deviate from its expected value [@Kaye1999Clarifying-the-]. Second, if the threshold is 99%, *at most* 1% of decision against defendants are expected to be mistaken (false convictions) and *at most* 99% of the decisions in favor of the defendant are expected to be mistaken (false acquittals). The exact ratio will depend on the probabilities assigned to defendants and how they are distributed [@allen2014]. The (expected) rate of false positives and false negatives—and thus their ratio—depend on where the threshold is located but also on the distribution of the liability probability as given by the density function $f(x)$.

Interval thresholds

The prior probability cannot be easily determined [@friedman2000}. Even if it can be determined, arriving at a posterior probability might be impractical because of lack of adequate quantitative information. Perhaps, decision thresholds should not rely on a unique posterior probability but on an interval of admissible probabilities given the evidence [@finkelstein1970bayesian]. %Perhaps, the assessment of the posterior probability of guilt can be viewed as an idealized process, a regulative ideal which can improve the precision of legal reasoning. (CITE BIEDERMAN TARONI).

Likelihood ratio thresholds

Add here stuff from Marcello's Mind paper about the prisoner hypothetical. Then, discuss Rafal's critique of the likelihood ratio threshold and see where we end up.

Completeness of the evidence

Discuss here Nance proposal. [@nance2016] argues that the evidence on which to base a trial decision should be reasonably complete—it should be all the evidence that one would reasonably expect to see from a conscientious investigation of the facts. A similar argument can be found in [@davidsonpargetter1987]. Arguably, probability-based decision thresholds can accommodate these considerations, for example, by lowering the probability of civil or criminal liability whenever the body of evidence is one-sided or incomplete [@Kaye79gate, @Kaye1986Do, @friedman1996]. Another strategy is to give a probability-based account of the notion of completeness of the evidence and other seemingly non-probabilistic criteria [@Urbaniak2017Narration-in-ju].

Risk of error threshold

Against numerical thresholds

When appellate courts have examined the question whether standards of proof can be quantified using probabilities, they have often answered in the negative. One of the clearest opposition to quantification was formulated by Germany's Supreme Court, the Federal Court of Justice, in the case of Anna Anderson who claimed to be a descendant of the Tsar family. In 1967, the Regional Court of Hamburg ruled

[2]The proof is the same as before. Since $tf(t) = 9(1-t)f(t)$, it follows that $t = 9/10$.

that Anderson failed to present sufficient evidence to establish that she was Grand Duchess Anastasia Nikolayevna, the youngest daughter of Tsar Nicholas II, who allegedly escaped the murder of the Tsar family by the Bolsheviks in 1918. %(Incidentally, DNA testing later demonstrated that Anna Anderson had no relationship with the Tsar family.) Anderson appealed to Germany's Federal Court, complaining that the Regional Court had set too demanding a proof standard. Siding with the lower court, the Federal Court made clear that `[t]he law does not presuppose a belief free of all doubts', thus recognizing the inevitable fallibility of trial decisions. The Court warned, however, that it would be`wrong' to think that a trial decision could rest on 'a probability bordering on certainty' (Federal Court of Justice, February 17, 1970; III ZR 139/67). This decision is all the more remarkable as it applies to a civil case.

% %but then warned that `this is often expressed imprecisely in such a way that the court may be satisfied with a probability bordering on certainty' and unequivocally concluded`this is wrong' . % Compared to civil cases, the resistance toward quantification can be more easily made plausible in criminal cases. [@buchak2014belief], for example, notes that an attribution of criminal culpability is an ascription of blame and such an ascription should require a full belief in someone's guilt, not just a probabilistic belief, however strong. One is left wondering, however. If a high probability of guilt short of 100% isn't enough but absolute certainty cannot be required either, how else could the standard of proof be met? The question becomes more pressing in civil cases. Anticipating this sort of worry, Germany's Federal Court in the Anderson case endorsed a conception of proof standards that echoed how U.S. courts describe proof beyond a reasonable doubt (see earlier in). The Federal Court wrote that a judge's decision must satisfy 'a degree of certainty which is useful for practical life and which makes the doubts silent without completely excluding them' (Federal Court of Justice, February 17, 1970; III ZR 139/67).

Leaving case law aside, there exist several theoretical alternatives to the probabilistic interpretation of proof standards in the scholarly literature. %Some scholars, on empirical or normative grounds, resist the claim that the point of gathering and assessing evidence at trial is solely to estimate the probability of the defendant's civil or criminal liability. [@Pennington1991, @penn1993] have proposed the *story model* according to which judges and jurors, first make sense of the evidence by constructing stories of what happened, and then select the best story on the basis of multiple criteria, such as coherence, fit with the evidence and completeness. % Along similar lines, [@Pardo2008Judicial-Proof-] argue that the version of the facts that best explains the evidence should prevail in a court of law. For a discussion of inference to the best explanation in legal reasoning, see [@schwartz2019WhatRelativePlausibility, @hastie2019CaseRelativePlausibilitya, @lai2019HowPlausibleRelative, @nance2019LimitationsRelativePlausibility]. % Another approach is due to [@gordon2007] and [@prakken2009] who view the trial as a place in which arguments and counterarguments confront one another. The party that has the best arguments, all things considered, should prevail. On this view, probability estimates can themselves be the target of objections and counterarguments. [@ho2008philosophy] and [@Haack2014-HAAEMS] hold that degrees of epistemic warrant for a claim, which depend on multiple factors – such as the extent to which the evidence supports the claim and it is comprehensive – cannot be equated to probabilities. [@stein2008] argues that, in order to warrant a verdict against the defendant, the evidence should have withstood objections and counterarguments, not merely supporting a high probability. [@gardiner2019ppa] argues that standards of proof should rule out all error possibilities that are relevant and these need not coincide with error possibilities that are probable. Finally, some epistemologists argue that a probabilistic belief, no matter how high, is not enough to warrant knowledge, and knowledge should be the standard for trial verdicts (see references in the next section).

There are a plethora of other objections. The puzzles of naked statistical evidence and the conjunction paradox are two of the most widely debated in the literature. These and other objections are examined in the sections that follow.