Is Proof Beyond a Reasonable Doubt a High Probability Threshold?

7,623 words / 48,392 characters – November 25, 2019

Abstract Some hold that the standard of proof in criminal cases—known as 'proof beyond a reasonable doubt'—is met whenever the probability of the defendant's guilt, given the evidence presented at trial, is sufficiently high, say, 95%. This proposal is plausible in many ways, but a number of objections have been leveled against it. Hypothetical cases exist in which the probability of guilt is clearly very high, yet people are unwilling to convict. Another objection is that the standard 'proof beyond a reasonable doubt' requires that guilt be established with reasonable certainty, yet a probability threshold cannot satisfy this requirement. These objections, are not decisive and the literature remains divided. To make progress, the paper focuses on a core function of the standard of proof, that is, to protect innocent defendants against mistaken convictions. As it turns out, the probability threshold theory of the standard of proof cannot accommodate this core function. To remedy this, the paper argues that a probability threshold should be supplemented with an additional requirement—roughly, that the evidence undergo an effective adversarial scrutiny at trial. This addition yields a better theory that can handle the main objections against the probability threshold theory.

1 Introduction

In a criminal case, the prosecutor must establish the defendant's guilt beyond a reasonable doubt, or else the defendant must be acquitted. This is the standard of proof that governs decisions in criminal trials. Some hold that this standard is met whenever the probability of the defendant's guilt, given the evidence presented at trial, reaches a sufficiently high threshold, say, 95%.[1] Call this the probability threshold interpretation or, more generally, the threshold theory. This interpretation is, in many respects, plausible. Given that certainty cannot be achieved, if a guilt probability of 95% is not enough to convict, what else is? This paper, however, first argues that the threshold theory of the standard of proof is problematic in some fundamental way and then puts forward a better theory.

I begin with examining two common objections against the threshold theory. One objection arises from the paradoxes of naked statistical evidence. These are hypothetical cases in which

[1] The probability of the defendant's guilt should be understood as the degree of belief that a rational agent who considered all the evidence would assign to the proposition 'the defendant is guilty.'

the probability of guilt is clearly very high, yet people are unwilling to convict on this ground alone. The second objection is that the standard 'proof beyond a reasonable doubt' demands that guilt be established with reasonable certainty, and a probability threshold, no matter how high, cannot meet this demand for certainty. These objections are intuitively compelling, but as Section 2 makes clear, they leave the defenders of the threshold theory largely unmoved. The debate in the literature, as I see it, is at a stalemate.

The paper attempts to break the stalemate by identifying a core function of the standard 'proof beyond a reasonable doubt' that any theory should be able to accommodate. This core function is to protect innocent defendants by reducing their risk of conviction. At first, it would seem that a demanding probability threshold should serve this function well. Section 3, however, shows that this is not the case. The threshold theory lacks the theoretical resources to explain how proof beyond a reasonable doubt protects the innocents against being exposed to an excessive risk of mistaken conviction. This is a shortcoming that cannot be taken lightly.

The paper puts forward a revised threshold theory that consists of two requirements: (1) a threshold probability requirement; and (2) an adversarial requirement on the evidence. Section 4 shows that, when these two requirements are combined, they do protect the innocents by reducing their risk of conviction. Requirement (2) expresses the simple idea that a defendant cannot be convicted unless the defense was able to effectively scrutinize the incriminating evidence. This is a familiar idea common to all modern legal systems. Yet its significance for understanding the standard of proof has been little discussed in the philosophical literature.

My hope is that the revised threshold theory will be favorably received by both supporters and critics of the original threshold theory. The supporters should welcome the revisions because they do not completely overhaul the original theory. The critics of the threshold theory should take the revisions seriously. As Section 5 and Section 6 show, the revised threshold theory can handle the two key objections against the original threshold theory. If there are intuitive reasons to resist the original theory, these no longer apply to its revised version.

2 THE DEBATE

The threshold interpretation of proof beyond a reasonable doubt has quite a few supporters. One of the early theorists of probability discussed whether the basis for a criminal conviction should be a guilt probability of 99% or whether the more demanding threshold of 99.9% should be required.[2] Statistical decision theory can help us to locate the appropriate probability threshold. For example, if the costs of a mistaken conviction are regarded as nine times the costs of a mistaken acquittal, a simple expected utility model would recommend that the probability threshold be set to 90%.[3] Some have argued that the probability threshold should be lower, say, 75%.[4] Others have argued that the threshold should vary depending on the costs at stake in a particular case or type of cases.[5] But despite these variations, many agree that a probability threshold in criminal cases must be located at a high or very high value.

Its popularity in some circles notwithstanding, the probability threshold interpretation counts many detractors. A recurrent worry is that it is far-fetched to quantify the probability of guilt since too many variables are at play. But although this worry is certainly justified, its force should not be exaggerated. In the interest of theorizing about the standard of proof, some abstracting away from the reality of trial proceedings can be tolerated. If it is unrealistic to effectively quantify the probability of the defendant's guilt, the quantification can be thought of as an idealization.

But even setting aside the practical difficulties of quantifying the guilt probability, other reservations of a theoretical nature remain. One line of criticism that is often pursued in the legal and philosophical literature relies on hypothetical scenarios. Here is an example. Suppose one hundred, identically dressed prisoners are out in a yard during recreation time. Suddenly, ninety-nine of them assault and kill the only guard on duty. We know that this is what happened from a video recording, but we do not know the identity of the ninety-nine killers. After the fact, a prisoner is picked at random and tried. Since he is one of the prisoners who were in the yard, the probability of his guilt would be 99%. But despite the high probability, many

[2] Bernoulli (1713).
[3] Laplace (1814); Kaplan (1968); Kaye (1999); Tillers and Gottfried (2007).
[4] Laudan (2011); for a critique, see Gardiner (2017).
[5] Kaplow (2012); for a critique, see Allen and Stein (2013); Picinali (2013).

have the intuition that this is not enough to establish guilt beyond a reasonable doubt.[6]

Another line of criticism points out that proof beyond a reasonable doubt is often understood as demanding certainty of guilt—not theoretical but practical, humanly attainable certainty, of the kind we expect when making decisions about the most important affairs of our lives.[7] This suggests that proof beyond a reasonable doubt requires a state of mind of being fully convinced about the defendant's guilt. Meeting a probability threshold, no matter how high, falls short of achieving such full conviction. Further—so the criticism goes—if proof beyond a reasonable doubt demands certainty, it should strive to eliminate any error without settling, in advance, for an acceptable margin of error. But a threshold that required less than certainty would fail to meet this demand because it would subject the innocents, deliberately, to the possibility of a mistaken conviction.[8]

Both criticisms have been met with some resistance. The first criticism heavily relies on the intuitive judgment that the high probability of the prisoners's guilt in the hypothetical scenario does not amount to proof beyond a reasonable doubt. But suppose we changed the numbers and imagined there were one thousand prisoners of whom nine hundred and ninety-nine killed the guard, so that the guilt probability of a prisoner picked at random would be 99.9%. Some have argued that when the guilt probability reaches such extremely high values, people's intuitive resistance to convicting should subside.[9] Another problem, as some have noted, is that intuitions in hypothetical scenarios are far removed from real cases and thus are potentially unreliable as a guide to theorize about the standard of proof.[10]

There is also resistance toward the second criticism of the threshold theory. This criticism alleges that by setting a decision threshold short of certainty, the trial system would deliberately—and thus wrongly—expose innocent defendants to the possibility of a mistaken conviction. But if guilt cannot be realistically established with certainty, aiming to a humanly attainable certainty would be no different than settling for an appropriately high threshold, or else convicting would become impossible. Unless the criticism is an invitation to convict no

[6] See Nesson (1979); Wells (1992); Redmayne (2008); Ho (2008); see also references in notes 12 through 15.

[7] Proof beyond a reasonable doubt is often paraphrased with 'moral certainty' or 'abiding conviction.'

[8] As Tribe (1971) puts it, 'formulating an "acceptable" risk of error to which the trier is willing deliberately to subject the defendant would interfere seriously with ... the demand for certitude' (p. 1347); see also Stein (2005).

[9] Roth (2010).

[10] Lempert (1986); Allen and Leiter (2001).

one, there seems to be no theory of the standard of proof that can meet its demands.[11]

On balance, while the two criticisms of the threshold theory have strong intuitive appeal, they are far from being decisive arguments against the theory. The defenders of the threshold theory remain largely unmoved. Nevertheless, people's intuitions in hypothetical cases such as the prisoner scenario are quite compelling. This has prompted many scholars to formulate theories that can vindicate these intuitions. Some have argued that other requirements must be added to, or replace, an appropriately high probability threshold as decision criteria for a conviction. Some of the additional or alternative requirements that have been proposed focus on properties of the incriminating evidence, for example, that it should be causally connected;[12] modally robust;[13] normically justifying;[14] sufficient for knowledge of guilt.[15]

Incorporating these requirements in a theory of the standard of proof would vindicate the intuition that an appropriately high probability of guilt, by itself, is not enough to establish guilt beyond a reasonable doubt.[16] And yet, a pressing question remains. The proposed additional or alternative requirements are rooted in theoretical considerations about the nature of epistemic warrant, but their relevance to the criminal trial is not immediately apparent. The defenders of the threshold theory can legitimately ask why, exactly, new requirements should be adopted as part of the standard of proof. Besides vindicating some contested intuitions, the proponents of the new requirements do not have a compelling answer here.[17]

To make progress, it might seem promising to ground a theory of the standard of proof on an overarching theory of judicial fact-finding or a theory of the trial. Some scholars, for example, have argued that evidential reasoning in the law should aim to offer an explanation of the evidence, not so much assess the probability of the defendant's guilt. The probability threshold theory, then, would be inadequate because it is mismatched with the nature of evidential reasoning in the law.[18] Yet, in response, the defenders of the threshold interpretation

[11] Walen (2015).
[12] Thomson (1986).
[13] Modally robust features are sensitivity and safety; see Enoch et al. (2012); Pritchard (2015).
[14] Smith (2018)
[15] Moss (2018).
[16] See, however, Blome-Tillmann (2015).
[17] Enoch et al. (2012) and Pritchard (2015) do discuss the legal relevance of their accounts, but as Smith (2018) points out, modal robustness is an externalist condition that can hardly be part of the standard of proof.
[18] Pardo and Allen (2008). For other theories of judicial fact-finding, see Cohen (1981); Stein (2005); Ho (2008); Picinali (2012); Nance (2016). For a theory of the trial, see Duff et al. (2007).

can invoke statistical decision theory and the maximization of expected utility as the overarching framework that justifies their own interpretation of the standard of proof. This would bring the discussion to another stalemate.[19]

The current debate, then, is caught between two extremes: an appeal to high-level theoretical frameworks or an appeal to intuitions about what counts as proof beyond a reasonable doubt. Since both can be hotly contested, the chances of a resolution are slim. A way out is to carve a middle way. The next section establishes some common ground by identifying a core function of the standard of proof in criminal cases. This can place the debate on a firmer footing. Once we agree on a core function of the standard of proof, we should be better equipped to evaluate the threshold theory in terms of how well it accommodates this core function.

3 Protecting the Innocents

There can be little doubt that avoiding mistaken convictions is of the utmost importance for the criminal trial, especially in light of the great harm that would result from convicting an innocent. This, of course, cannot mean that innocent defendants should never be convicted. Given the fallibility of the trial system, innocent defendants are exposed to a risk of mistaken conviction. This risk is determined, to a large extent, by the procedural protections that are afforded to defendants, such as the presumption of innocence, the right to a defense, the standard of proof, etc. Together with other trial protections, then, the standard 'proof beyond a reasonable doubt' is expected to limit the risk of mistaken conviction. As the United States Supreme Court put it, proof beyond a reasonable doubt is 'a prime instrument for reducing the risk of convictions resting on factual error.'[20]

But how, exactly, should the risk of mistaken conviction be understood? And to what extent is the standard of proof expected to reduce the magnitude of this risk? Let's begin with the first question. One possibility is to view the risk of mistaken conviction simply as a function of the probability threshold. If the threshold is 99%, the risk of mistaken conviction would be at most 1%; if the threshold is 95%, the risk would be at most 5%; and so on. The thought is that if the probability threshold for conviction is set at 99%, a convicted defendant

[19] The debate between **?** and **?**, though illuminating, has this flavor.
[20] *re Winship* (1970), 397 U.S. 358, 363.

must be assigned a guilt probability of 99% or higher, and this means that there is at most a 1% probability that the defendant is innocent. The same holds *mutatis mutandis* for a probability threshold of 95%, and similarly for other thresholds.

Note that the upper bounds 1%, 5% and the like—values that can be read off the probability threshold—express the conditional probability that if a defendant is convicted (abbreviated C), this defendant is innocent (abbreviated I_{def}), or in symbols, $Pr(I_{def}|C)$.[21] The proposal, then, would be that the risk of mistaken conviction for innocent defendants should be understood in terms of this conditional probability, so that the lower this probability, the better the protection against mistaken convictions. In this framework, setting the appropriate value for the risk of mistaken conviction would be the same as setting the appropriate value for the probability threshold.[22] But things are not as simple as they seem. I will now show that understanding the risk of mistaken conviction in terms of the conditional probability $Pr(I_{def}|C)$ is misguided.

The risk we want to capture should reflect the extent to which the trial system—and proof beyond a reasonable doubt in particular—protects innocent defendants against mistaken convictions. We expect that the stronger the protection, the lower the risk. We also expect that when the protection is entirely missing, the risk would be at its maximum. But these expectations are not met if we understand the risk of mistaken conviction in terms of the conditional probability $Pr(I_{def}|C)$ that if a defendant is convicted, this defendant is innocent. For consider a trial system that is characterized by the following feature: 99% of its defendants are guilty (hence, 1% are innocent), and the rule of decision is to convict everyone who faces trial no matter what. Intuitively, this rule of decision would offer no protection to the innocents facing trial. We would therefore expect the risk of mistaken conviction to be 100%. Yet since by assumption 1% of the defendants are innocent, the conditional probability $Pr(I_{def}|C)$ that a defendant, if convicted, is innocent would be 1%. The defendants in this trial system would seem to enjoy some protection—even a strong one since 1% is a low value—but this cannot be

[21] If the trial is a binary classifier, this probability can be approximated by the long-run ratio between the number of factually innocent defendants who haven been incorrectly classified as guilty (i.e. convicted) divided by the number of defendants, factually guilty or innocent, who have been, correctly or incorrectly, classified as guilty (i.e. convicted).

[22] Stein (2005) relies on a similar conception. He argues that the proof standard in criminal cases should fully immunize defendants against the (evidenced) risk of mistaken conviction. This risk is greater than zero whenever, based on evidence presented at trial, there remains a possibility that the defendant facing trial is innocent. If this risk is greater than zero, Stein believes that the defendant should be acquitted, a position known in the literature as maximalism; see footnote 32.

right because defendants are in fact given no protection at all.

To avoid this counterintuitive result, the risk of mistaken conviction should be understood as the conditional probability that a defendant, if innocent, is mistakenly convicted, or in symbols, $Pr(C|I_{def})$. This can be roughly approximated—over a long period of time—by the ratio between the number of innocents who are convicted and the total number of innocents facing trial. This conditional probability plausibly captures the extent to which the trial system protects innocent defendants against mistaken convictions. Regardless of the percentage of innocents among the defendants facing trial, the conditional probability $Pr(C|I_{def})$ would still be 100% so long as the rule of decision was to convict every defendant no matter what. This is what we expect because in such a trial system innocent defendants would have no protection.

If the conditional probability $Pr(C|I_{def})$ is the right way to understand the risk of mistaken conviction, the next question is to what extent the standard 'proof beyond a reasonable doubt' should reduce the magnitude of this risk. I will tackle this question in the next section. For now I will rely on the modest claim that the standard of proof should reduce the risk of mistaken conviction to some extent, that is, it should keep it below 100%. This might seem a trivial requirement, but it is not. We will see that even a demanding probability threshold as the rule of decision can do little or nothing to keep $Pr(C|I_{def})$ below 100%.

Suppose a trial system enforces a demanding probability threshold for conviction, but gives defendants no opportunity to scrutinize the evidence against them or present exculpatory evidence. In this trial system—as in any other—prosecutors want to maximize their chances of winning a case, and thus they will send to trial only those suspects against whom there is strong incriminating evidence that can meet the probability threshold for conviction. What is peculiar about this trial system is that since defendants are not allowed to scrutinize the evidence against them, prosecutors will typically win their cases. In light of there being strongly incriminating evidence that is unchallenged, the decision-makers will assign defendants a high probability of guilt, exceeding the threshold for conviction. Nearly all defendants, including innocent defendants, will be convicted, and thus the risk of mistaken conviction—that is, the conditional probability $Pr(C|I_{def})$—will approximate 100%.[23]

[23] Prosecutors might be unable to predict, exactly, what jurors would decide. So $Pr(C|I_{def})$ might be slightly below 100%, but since this would be a random effect, the probability threshold would play no systematic role in keeping $Pr(C|I_{def})$ below 100%.

Some might be puzzled by this result. The risk of mistaken conviction equals 100% under a rule of decision that convicts every defendant no matter what. This is to be expected. But if the above argument is correct, this risk still approximates 100% under a more discerning rule of decision. How can that be? It is puzzling that applying a demanding threshold would not reduce, in any systematic way, the risk of mistaken conviction compared to a rule that convicts everyone no matter what. After all, it seems perfectly plausible to argue, as some have done, that if the threshold is made more stringent, the risk of conviction for both innocent and guilty defendants should diminish to some extent.[24] But this would be a mistake.

We should keep in mind that the risk of mistaken conviction—that is, the probability $Pr(C|I_{def})$ that a defendant, if innocent, is convicted—measures the degree of protection for innocent defendants at trial. The defendants facing trial are selected from a larger pool of suspects. This mechanism of selection is driven by prosecutorial choices and depends heavily on what we might call *feedback mechanisms*. In particular, if prosecutors know that jurors would apply a certain threshold at trial, they would bring to trial only those suspects who confront evidence that could reasonably meet the threshold applied by jurors. After all, why send to trial those who have little or no chance of being convicted? When the rule of decision is to convict everyone who faces trial, prosecutors are under no pressure to make a selection since any suspect they send to trial will be convicted no matter what.

But now suppose the threshold of decision is made more stringent, say, it is set at 95% probability, while defendants still have no right to present exculpatory evidence. In this case, since they can no longer be sure that whoever they send to trial will be convicted, prosecutors will be under some pressure to make a selection. Presumably, they will send to trial only those suspects who face incriminating evidence that could reasonably meet the 95% probability threshold. But since all defendants who end up at trial will face incriminating evidence that can reasonably meet the more stringent threshold, their risk of conviction—as defendants—will still approximate 100%. Not much would change if the probability threshold were more stringent, say, it was set at 99%, while defendants still had no right to present exculpatory evidence. Prosecutors would send to trial only those suspects facing incriminating evidence that could reasonably meet the 99% threshold. Once at trial, this more selective pool of defendants

[24] See, among others, (Dekay, 1996) who uses the framework of Signal Detection Theory to make this point.

would still be exposed to a risk of mistaken conviction approximating 100%.

No matter how stringent, a probability threshold—as the sole rule of decision—cannot reduce the risk of mistaken conviction for innocent defendants in any systematic way. This should put pressure on the defenders of the threshold theory. If proof beyond a reasonable doubt is meant to protect innocent defendants against mistaken convictions, at least to some extent, the threshold theory fails to accommodate this function of the standard of proof.

Two responses on behalf of the threshold theorist are worth discussing. First, even if a stringent threshold can let the risk of mistaken conviction for innocent defendants surge to 100%, such a threshold does still limit the risk of mistaken conviction to some extent. A stringent threshold—because of feedback mechanisms—will force prosecutors to make a selection among the suspects to send to trial. By reducing the risk that innocent suspects would face trial, a stringent threshold reduces the risk of mistaken conviction for them.

Here is a counter to this response. The standard of proof operates as part of the trial, and as such, it is expected to offer protection to those who face trial, namely defendants. That it protects, indirectly, innocent suspects is a welcome addition, but cannot be its sole function. The trial system should offer to innocent *defendants* an extra layer of protection on top of the protection afforded by pre-trial proceedings to innocent suspects. When innocent citizens become suspects who could face trial, they should enjoy protection against mistaken charging decisions, and when suspects become defendants facing trial, they should enjoy protection against mistaken convictions. As part of the trial, the standard of proof should reduce the risk of mistaken conviction for defendants, not just the risk for suspects or citizens at large.

And now consider a second response on behalf of the threshold theorist. Note that the examples above—which showed how a stringent probability threshold, no matter how stringent, would still let the risk for innocent defendants surge to 100%—all depicted a trial system that gave defendants no opportunity to respond to the evidence presented against them. The point of these examples is clear enough. That is, if the prosecutor was the only party who could present evidence, and the defense had no saying, innocent defendants would enjoy little or no protection at trial even when the rule of decision required that guilt be established with an extremely high probability. But does this really pose a problem for the threshold theory? Since any modern legal system gives the defense an opportunity to present its own evidence, this

cannot be an objection against the threshold theory *per se*. It would be an objection against an unduly restrictive reading of the theory, what might be called the bare threshold theory according to which trial decisions are made without taking into account evidence from both sides. On a more charitable reading, the threshold theory should be interpreted as issuing the following decision rule: convict only if the guilt probability is above some threshold, where the guilt probability is estimated on the basis of all the evidence presented at trial, both incriminating and exculpatory. There is no reason for thinking that under such conditions the risk of mistaken conviction would be 100%.

This response concedes that the bare threshold theory is untenable. If the evidence is allowed to be one-sided, innocent defendants will enjoy no protection at trial, even when guilt is established with an extremely high probability. The untenability of the bare threshold theory shows that the body of evidence used to assess the guilt probability must satisfy some other requirement. But what is, exactly, this other requirement? The suggestion that the body of evidence should contain all the evidence, both incriminating and exculpatory, is promising, but a more principled discussion is necessary. This is the task of the next section.

4 The Revised Threshold theory

It is sensible to expect that the evidence presented at trial should be both incriminating and exculpatory. In many trial systems, defendants enjoy a right to a defense, for example, they have a right to cross-examine the witnesses who testify against them. In this legal framework whose details vary from country to country, the evidence presented at trial is the product of an adversarial process in which the prosecution makes its case and the defense responds to the best of its ability. With this in mind, a natural way to revise the threshold theory is as follows: the defendant's guilt is established beyond a reasonable doubt whenever (1) the probability of guilt, based on all the evidence presented at trial, reaches an appropriately high threshold, where (2) the evidence presented at trial is the result of a process in which the defense had an opportunity to respond to the prosecutor's case and took full advantage of this opportunity, and the prosecution in turn also had an opportunity to respond to the challenges posed by the defense and took full advantage of this opportunity.

Condition (1), by itself, is the threshold interpretation of 'proof beyond a reasonable doubt.' The additional condition (2) is a specification of condition (1). It fits naturally in the adversarial structure of judicial fact-finding and thus its relevance for trial proceedings is immediately apparent. But before adding it to our theory of the standard of proof, we should clarify what it is expected to accomplish. In particular, does condition (2)—in combination with a high probability threshold as in condition (1)—ensure that innocent defendants are protected against mistaken convictions? Just as a high threshold probability can let the risk of mistaken conviction surge to 100%, so can a merely perfunctory, shallow and ineffective scrutiny of the incriminating evidence. To keep the risk below 100%, the adversarial scrutiny of the evidence must be effective. But what does this require, exactly? I will argue that the process of adversarial scrutiny is effective if if satisfies two normative requirements: diagnosticity and best effort diagnosticity. I will now discuss each in turn.

Suppose that, at the very beginning of the trial, the prosecutor presents a strong incriminating case against the defendant. If the fact-finders were to consider only the prosecutor's evidence, they would assign the defendant a very high guilt probability, exceeding the threshold for conviction. Whether innocent or guilty, the defendant would presumably be convicted, and thus the risk of conviction would approximate 100%. The process of adversarial scrutiny, however, allows the defense to challenge the incriminating evidence. This adversarial scrutiny should allow some innocent defendants to bring their guilt probability below the threshold and escape conviction. Their risk of mistaken conviction should thus be below 100%.

Yet the fact that innocent defendants are subject to a risk of conviction that is below 100% is not quite enough. For suppose that once a defendant faced trial, the fact-finders tossed a weighted coin having, say, a 1% probability of landing heads. Suppose the rule of decision was that if the coin landed heads, they would convict, and otherwise, they would acquit. Although this rule should be kept hidden from the public to preserve the system's credibility, it would be effective at protecting innocent defendants. Their risk of conviction would only be 1%, and it could become even lower by decreasing the probability of 'heads.' So why bother with the process of adversarial scrutiny if tossing a coin can protect defendants so effectively?

The answer is that, while tossing a weighted coin can ensure that the risk of mistaken conviction is only 1%, or whatever suitably low value we desire, the risk of correct conviction

for guilty defendants will also be as low as 1%. If trial decisions were taken by tossing a coin, their accuracy would be extremely poor because both innocents and guilty defendants would be subject to same risk of conviction. The trial system should instead ensure that, once they face trial, innocent defendants are subject to a lower risk of conviction than guilty defendants. Call this property *diagnosticity*.[25]

Clearly, a rule of decision that solely consisted of a probability threshold would fail diagnosticity. As long as guilty and innocent defendants are assigned the same guilt probability, a probability threshold applies indiscriminately to both of them. But would the additional condition (2), in combination with condition (1), satisfy diagnosticity? As the process of scrutiny unfolds with inputs from both defense and prosecution, the incriminating evidence should normally be stronger for guilty than for innocent defendants, and the exculpatory evidence should normally be stronger for innocent than guilty defendants, other things being equal.[26] Hence, it should be easier for an innocent than a guilty defendant to bring the guilt probability below the threshold and escape conviction. This means that the risk of conviction should normally be lower for an innocent than for a guilty defendant. If this is correct, condition (2), in combination with (1), does satisfy diagnosticity.

This is all well and good, but diagnosticity does not seem enough either. If the risk of mistaken conviction for an innocent defendant is slightly less than the risk of correct conviction for a guilty defendant, say, 99% v. 99.1%, this would not justify the expenditure in efforts and resources that a trial requires. Would 40% v. 80% suffice? Why not require 1% v. 99% or 0.01% v. 99.99%? A plausible answer to these questions is that the risks of conviction should reflect the best efforts of both parties at trial. There are, of course, limits to these efforts, in terms of time and resources available. But within these limits, no risk of mistaken conviction should be considered, in principle, acceptable—not 5%, not 1%, not even 0.1%—if it can be reduced further at acceptable costs. By the same token, no risk of correct conviction should be considered enough—not 95%, not 99%, not even 99.9%—if it can be increased further at acceptable costs. The trial system, then, should ensure that the risk of mistaken conviction for innocent defen-

[25] More precisely, $Pr(C|I_{def}) < Pr(C|G_{deg})$, where I_{def} and G_{def} stand for 'the defendant is factually innocent' and 'the defendant is factually guilty'.

[26] The qualification 'other things being equal' is important because the exculpatory evidence in favor of a wealthy guilty defendant will probably be stronger than the exculpatory evidence in favor of a poor innocent defendant.

dants comes as close to 0% as possible, and the risk of correct conviction for guilty defendants as close to 100% as possible. Call this property *best effort diagnosticity*.

How can the trial system promote compliance with best effort diagnosticity? Condition (2) is of service here. During the adversarial scrutiny at trial, prosecution and defense will present evidence and arguments in support of their respective positions. We normally expect evidence and arguments to track, at least to some extent, factual guilt and factual innocence. We also expect that the more careful, extensive, thorough the adversarial scrutiny, the more likely it is that innocent defendants will bring their guilt probability below the threshold and escape conviction, and the less likely it is that guilty defendants will bring their guilt probability below the threshold. This means that the more extensive the scrutiny of the evidence from both sides, the lower the risk of mistaken conviction and the higher the risk of correct conviction. Crucially, condition (2) does not set a fixed limit to how extensive the scrutiny of the evidence should be. The expectation is that both prosecution and defense do their best, consistently with the existing constraints in time and resources. A scrutiny of the evidence that is as extensive as reasonably possible, then, should promote a cost-justified minimization of the risk of mistaken conviction and a cost-justified maximization of the risk of correct conviction. This should satisfy best effort diagnosticity.

Let me summarize. I have argued that, in combination with a threshold condition, the adversarial condition in the revised threshold theory would normally satisfy the normative requirements of diagnosticity and best effort diagnosticity. This rests on the plausible assumption that the evidence presented at trial normally tracks facts about guilt and innocence.[27] The upshot is that whenever the adversarial scrutiny of the evidence fails to satisfy either of these requirements, the scrutiny does not count as effective and thus cannot be the basis for a conviction. As we will now see, this helps to make progress toward addressing the two main objections against the threshold theory. The lack of diagnosticity in the decision process explains why a conviction in the prisoner scenario should be resisted even though the guilt probability is extremely high (§5). Further, best effort diagnosticity helps to make sense of the demand for practical certainty that underpins proof beyond a reasonable doubt (§6).

[27] This remains, of course, an assumption. I examine this assumption more in detail in other work.

5 BACK TO THE PRISONER SCENARIO

The puzzle in the prisoner scenario is that even though the guilt probability of the prisoner picked at random is high, this is—intuitively—insufficient to establish guilt beyond a reasonable doubt. The challenge is to justify this intuition in terms of the revised threshold theory, comprising conditions (1) and (2). At first blush, this would seem easy to do. While the probability of the prisoners's guilt is high, this is only the prosecutor's side of the story. The prisoner scenario, as originally described, does not say whether the incriminating evidence was scrutinized or whether an opportunity to scrutinize it was given to the defendant. If so, condition (1) would be satisfied, but (2) would not be.

This argument, however, is too quick. The prisoner on trial is not prohibited from scrutinizing the statistics brought against him. If other prisoners were interrogated, they could provide more details about what happened. The defense could also offer other statistics showing that the prisoner on trial participated, say, in only 1% of the prison riots in the past year.[28] If we included all the evidence that could be presented by both prosecutor and defense, whether the prisoner should be convicted or not will depend on the overall strength of the evidence presented. Why, then, would condition (2) not be met in the prisoner scenario?

To make progress here, we should step back and consider, in general, how a defendant can respond to an accusation based on some piece of evidence E. A defendant can respond by pursuing one of two routes: first, by eliciting additional information about E with the goal of *undercutting* the evidentiary connection between E and the proposition p that E purports to establish; and second, by offering independent evidence that *rebuts* the claim that p and supports a proposition incompatible with p.[29] For example, when an eyewitness makes a claim that incriminates a defendant, the undercutting route is typically open so long as it is possible to elicit additional information by asking questions to the witness. What were the lighting conditions? How far were you? And so on. These probing questions could prove the testimony unreliable. Crucially, this undercutting route would be open even if the rebutting route were closed because other independent evidence could not be found.

Consider now the prisoner scenario in light of these two defense routes. Interestingly, both

[28] For a discussion of this topic and the reference class problem, see Colyvan et al. (2001); Nance (2007).
[29] On this distinction, see Pollock (1987).

routes seem to be blocked. By stipulation, the scenario contains no other evidence besides the statistics. The rebutting route is therefore blocked. What about the undercutting route? The prisoner on trial can, of course, try to scrutinize the statistics and hope to prove them unreliable or untruthful. But unless there was a mistake in counting, nothing would change about the statistics after scrutinizing them in this way. Their scrutiny could hardly show them to be defective. So the undercutting route is also blocked.[30]

The prisoner scenario would seem to satisfy, on their face, conditions (1) and (2). After all, the guilt probability is high and the statistics are resistant against scrutiny because both routes of defense are blocked. This could suggest that the incriminating evidence is strong and that a conviction on its basis is justified. But is that so? In the prisoner scenario, the fact that both routes of defense are blocked does not depend, in any essential way, on whether the prisoner is guilty or innocent. Without access to other evidence, the probability of bringing the guilt probability below the threshold for conviction, upon scrutinizing the statistics, is equally slim, if not entirely inexistent, for a guilty and an innocent prisoner alike. This is a failure of what I called diagnosticity. On this reading, the paradoxical nature of the prisoner scenario is that both conditions (1) and (2) are met on their face, but on a deeper level—one that takes into account the diagnosticity of the process of adversarial scrutiny—condition (2) is not met.

It is instructive to draw a comparison with eyewitness testimony. If an eyewitness testified against the prisoner, the testimony would be considered enough, by itself, to make a *prima facie* case against the prisoner. If, in addition, the testimony withstood vigorous adversarial scrutiny in the form of, say, cross-examination, this should be enough to convict the prisoner. Could we say the same for the statistics in the prisoner scenario? Could we say that they make a *prima facie* case against the prisoner, since the guilt probability is so high, and so long as they can withstand scrutiny at trial, they should be enough to convict the prisoner?

The crucial point here is whether 'withstanding scrutiny' can be a reliable indicator of factual guilt. The scrutiny of eyewitness testimony is expected to be diagnostic. That is, when incriminating eyewitness testimony is scrutinized, the guilt probability based on that evidence should normally be brought below the threshold for conviction more often given

[30]Eyewitness testimony is interpretable. More information can be supplied during cross-examination and this information can strengthen or weaken the testimony. The statistics in the prisoner scenario are, this sense, less interpretable. On this point, see Allen (1991).

innocence than given guilt.[31] And when a piece of incriminating evidence withstands a *diagnostic* scrutiny, this fact can be taken as a reliable indicator of factual guilt. That is why eyewitness testimony can, at least in principle, be the basis for a conviction. The same does not apply to the statistics in the prisoner scenario. They can withstand scrutiny because both routes of defense are blocked given the stipulations in the scenario. But, as seen earlier, this process cannot be diagnostic in the sense that the probability that the prisoner can escape conviction—by scrutinizing the statistics and bringing the guilt probability below the threshold—is the same in the case of the prisoner's innocence or guilt. And when a piece of incriminating evidence can withstand a *non-diagnostic* scrutiny, this cannot be a reliable indicator of factual guilt. That is why the statistics in the prisoner scenario cannot be the basis for a conviction.

6 Practical certainty

Consider now the second criticism of the threshold theory. This criticism is rooted in the demand for practical certainty that underpins, at least intuitively, proof beyond a reasonable doubt. It is difficult, however, to make sense of practical certainty within the bare threshold theory. This theory leaves open only two options. Either we settle for some suitably high guilt probability threshold as the rule of decision, or we require that guilt be established with 100% probability. The former option reasserts the threshold theory and the latter makes convicting impossible. This section will show that the revised threshold theory can offer a plausible interpretation of practical certainty. In fact, I will offer two interpretations of this idea.

First, practical certainty can be understood in terms of cost-justified risk minimization. To illustrate, suppose a trial system imposes a 25% risk of conviction on its innocent defendants and as a consequence ends up convicting, say, 10,000 innocent people per month. The question is whether this system is morally blameworthy for the mistaken convictions. For one thing, the system should not be blamed simply because it convicted some innocent defendants. For another, the system should be blamed if it could have convicted fewer innocents by reducing the risk of mistaken conviction. It is instructive to contrast two cases here. First: additional evidence could have been found at affordable costs and this evidence, if presented at trial,

[31] Whether this is true is an empirical question. If it turned out to be false, we should reassess the value we place on the cross-examination of eyewitnesses.

would have halved the risk of mistaken conviction. In this case, the system should be blamed for the excess innocents who were convicted. Second: the risk could have been reduced further only by giving every defendant unlimited funds. Since this would be too costly, the trial system should not, in this case, be blamed for its mistaken convictions.

This is nothing new. As noted earlier while discussing best effort diagnosticity, no risk of error for the innocents should be considered, in principle, acceptable—not 5%, not 1%, not even 0.1%—if it could have been reduced further at acceptable costs. We have already seen why the revised threshold theory can comply with best effort diagnosticity. Condition (2) requires that the evidence be scrutinized as extensively as possible, but does not set a limit to how extensive the scrutiny should be. The more extensive the examination of the evidence, the lower the risk of mistaken conviction. By complying with best effort diagnosticity, the revised threshold theory can accommodate practical certainty, understood as the requirement that the risk of mistaken conviction be kept as close to 0% as (reasonably) possible.[32]

The revised threshold theory can accommodate another interpretation of practical certainty, one in terms of *full belief*. Intuitively, guilt is established beyond a reasonable doubt when the fact-finders at trial are fully convinced, in some suitable sense, about the defendant's guilt, and not simply when the defendant's guilt is highly probable.[33] Full belief, of course, cannot be construed as requiring that the defendant's guilt be 100% probable, or else convicting would become impossible. As some have suggested, full belief can be understood as highly probable belief that is probabilistically stable even in light of potentially countervailing evidence. More precisely, a belief in p counts as full belief if p is highly probable, given the evidence available, and would remain highly probable even considering other evidence.[34]

Note that the body of evidence relative to which the stability of the belief is assessed must be reasonably circumscribed, or else stable belief would be a belief that no evidence can make less probable. The revised threshold theory of the standard of proof can be of service here. Recall, the theory comprises two conditions: (1) the probability of the defendant's guilt should

[32]The view that the standard of proof should keep the risk of mistaken conviction as low as possible is sometimes called *maximalism*; see, among others, Duff et al. (2007) and Stein (2005). Maximalists, however, are not explicit about their view of the risk of mistaken convictions. They seem to aspire to the minimization of the risk that a defendant, if convicted, is innocent, not the minimization of the risk that the defendant, if innocent, is convicted. On this distinction, see Section 3. For a critical discussion of maximalism, see Walen (2015). .

[33]This might be another reason why we feel uneasy in convicting in the prisoner scenario; see Buchak (2013).

[34]See, among others, Skyrms (1980); Leitgeb (2014).

be appropriately high given the evidence; (2) the prosecution and the defense should be given the opportunity to examine the evidence and took advantage of this opportunity to the best of their ability. Condition (2) can be read as circumscribing the body of evidence relative to which the stability of the belief about guilt should be assessed. That is, the body of evidence in question comprises whatever information the defense brought up, along with the response by the prosecution. So whenever conditions (1) and (2) are satisfied, the resulting belief about the defendant's guilt will be a highly probable belief that is stable in the relevant sense. This is a characterization of full belief that does not equate it with infallible belief.

REFERENCES

Allen, R. J.: 1991, 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.

Allen, R. J. and Leiter, B.: 2001, Naturalized epistemology and the law of evidence, *Virginia Law Review* **87**, 1491–1550.

Allen, R. J. and Stein, A.: 2013, Evidence, probability and the burden of proof, *Arizona Law Journal* **55**, 557–602.

Bernoulli, J.: 1713, *Ars conjectandi*.

Blome-Tillmann, M.: 2015, Sensitivity, causality, and statistical evidence in courts of law, *Thought: A Journal of Philosophy* **4**(2), 102–112.

Buchak, L.: 2013, Belief, credence, and norms, *Philosophical Studies* **169**(2), 285–311.

Cohen, J. L.: 1981, Subjective probability and the paradox of the Gatecrasher, *Arizona State Law Journal* pp. 627–634.

Colyvan, M., Regan, H. M. and Ferson, S.: 2001, Is it a crime to belong to a reference class?, *Journal of Political Philosophy* **9**(2), 168–181.

Dekay, M. L.: 1996, The difference between Blackstone-like error ratios and probabilistic standards of proof, *Law and Social Inquiry* **21**, 95–132.

Duff, A., Farmer, L., Marshall, S. and Tadros, V.: 2007, *The Trial on Trial (vol 3): Towards a Normative Theory of the Criminal Trial*, Hart Publishing.

Enoch, D., Fisher, T. and Spectre, L.: 2012, Statistical evidence, sensitivity, and the legal value of knowledge, *Philosophy and Public Affairs* **40**(3), 197–224.

Gardiner, G.: 2017, In defence of reasonable doubt, *Journal of Applied Philosophy* **34**(2), 221–241.

Ho, H. L.: 2008, *Philosophy of Evidence Law*, Oxford University Press.

Kaplan, J.: 1968, Decision theory and the fact-finding process, *Stanford Law Review* **20**(6), 1065–1092.

Kaplow, L.: 2012, Burden of proof, *Yale Law Journal* **121**(4), 738–1013.

Kaye, D. H.: 1999, Clarifying the burden of persuasion: What Bayesian rules do and not do, *International Commentary on Evidence* **3**, 1–28.

Laplace, P.: 1814, *Essai philosophique sur les probabilités*.

Laudan, L.: 2011, The rules of trial, political morality and the costs of error: or, is proof beyond a reasonable doubt doing more harm than good?, *in* L. Green and B. Leiter (eds), *Oxford Studies in Philosophy of Law*, Vol. 1, Oxford University Press.

Leitgeb, H.: 2014, The stability theory of belief, *Philosophical Review* **123**(2), 131–171.

Lempert, R. O.: 1986, The new evidence scholarship: Analysing the process of proof, *Boston University Law Review* **66**, 439–477.

Moss, S.: 2018, *Probabilistic Knowledge*, Oxford University Press.

Nance, D. A.: 2007, The reference class problem and mathematical models of inference, *International Journal of Evidence and Proof* **11**(4), 259–273.

Nance, D. A.: 2016, *The Burdens of Proof: Discriminatory Power, Weight of Evidence, and Tenacity of Belief*, Cambridge University Press.

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

Pardo, M. S. and Allen, R. J.: 2008, Judicial proof and the best explanation, *Law and Philosophy* **27**(3), 223–268.

Picinali, F.: 2012, Structuring inferential reasoning in criminal fact finding: An analogical theory, *Law, Probability and Risk* **11**(2/3), 197–223.

Picinali, F.: 2013, Two meanings of "reasonableness": Dispelling the "floating" reasonable doubt, *Modern Law Review* **76**(5), 845–875.

Pollock, J. L.: 1987, Defeasible reasoning, *Cognitive Science* **11**, 481–518.

Pritchard, D.: 2015, Risk, *Metaphilosophy* **46**(3), 436–461.

Redmayne, M.: 2008, Exploring the proof paradoxes, *Legal Theory* **14**(4), 281–309.

Roth, A.: 2010, Safety in numbers? Deciding when DNA alone is enough to convict, *New York University Law Review* **85**(4), 1130–1185.

Skyrms, B.: 1980, *Causal Necessity*, Yale University Press.

Smith, M.: 2018, When does evidence suffice for conviction?, *Mind* **127**(508), 1193–1218.

Stein, A.: 2005, *Foundations of Evidence Law*, Oxford University Press.

Thomson, J. J.: 1986, Liability and individualized evidence, *Law and Contemporary Problems* **49**(3), 199–219.

Tillers, P. and Gottfried, J.: 2007, Case comment–United States v. Copeland, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): A collateral attack on the legal maxim that proof beyond a reasonable doubt is unquantifiable?, *Law, Probability and Risk* **5**(2), 135–157.

Tribe, L. H.: 1971, Trial by mathematics: Precision and ritual in the legal process, *Harvard Law Review* **84**(6), 1329–1393.

Walen, A.: 2015, Proof beyond a reasonable doubt: A balanced retributive account, *Louisiana Law Review* **76**(2), 355–446.

Wells, G. L.: 1992, Naked statistical evidence of liability: Is subjective probability enough?, *Journal of Personality and Social Psychology* **62**(5), 793–752.