



Meta-uncertainty and the proof paradoxes

Katie Steele¹  · Mark Colyvan²

Accepted: 15 March 2023 / Published online: 4 May 2023
© The Author(s) 2023

Abstract

Various real and imagined criminal law cases rest on “naked statistical evidence”. That is, they rest more or less entirely on a probability for guilt/liability derived from a single statistical model. The intuition is that there is something missing in these cases, high as the probability for guilt/liability may be, such that the relevant standard for legal proof is not met. Here we contribute to the considerable debate about how this intuition is best explained and what it teaches us about evidential reasoning in the legal setting. We part ways with the recent scholarship, however. Unlike most others, our diagnosis is not that there is an important qualitative property that some evidence or bodies of evidence have, and that naked statistical evidence, most strikingly, lacks. Rather, we see cases resting on naked statistical evidence as lying at the extreme of a continuum of cases that are vulnerable to challenge due to “meta-uncertainty” about the underlying model.

Keywords Legal evidence · Proof paradoxes · Meta-uncertainty · Legal probabilism · Statistical evidence

1 Introduction

There are a number of real legal cases that have rested almost entirely on a probability for guilt/liability derived from a single statistical model. These cases have resulted in differing verdicts, but never straightforward verdicts; all have provoked debate about legal standards of proof. One such case is Betty Smith’s civil lawsuit

✉ Katie Steele
katie.steele@anu.edu.au

Mark Colyvan
mark.colyvan@sydney.edu.au

¹ School of Philosophy, Australian National University, Canberra, ACT 2600, Australia

² Department of Philosophy, University of Sydney, A14 Main, Quadrangle, Sydney, NSW 2006, Australia

against Rapid Transit Inc. (*Smith v Rapid Transit Inc.*, 317 Mass. 469, 470, 58 N.E. 2d 754, 755 (1945)).¹ Betty Smith had a damaging encounter with a bus and, while she had no evidence about the appearance of the bus, she claimed that Rapid Transit should pay damages. Her case rested on the fact that Rapid Transit was the only bus company officially operating on the street where the accident occurred, and hence more likely than not to be responsible for the accident (allowing for the possibility that some other bus company may have happened to drive down the street at the time and caused the accident).² What is puzzling is that the probability that Rapid Transit was responsible (> 0.5) is apparently sufficiently high to meet the relevant legal standard of proof, yet many express reservations about a decision to award damages. The reservations seem to arise from misgivings about the reliance on a probability derived more or less exclusively from a single statistical model, i.e., on “naked statistical evidence”.³

The puzzle has been taken up by philosophers and legal scholars in what now amounts to a large literature. While many purport to simply *explain* common reactions to cases like *Smith v Rapid Transit Inc.*, the descriptive task is clearly thought significant due to its potential normative implications—that there is not just a *perceived* but also an *actual* deficiency with naked statistical evidence, as far as legal proof is concerned. (The normative message is even more obvious where the reactions to be explained are those of the expert reader, as opposed to those of the general populace.) We too will treat reactions to cases like *Smith v Rapid Transit Inc.* as flagging an interesting normative lesson. But our approach is cautious: there are, after all, many documented biases in our probabilistic reasoning.⁴

To be sure, it is because the familiar probabilistic reasoning biases do *not* seem to explain reactions to the classic cases of naked statistical evidence that the puzzle these cases present is so compelling. Indeed, many regard the probabilities in question to be uncontroversial, and think they clearly meet the threshold typically associated with the relevant standard of proof: $\geq 50\%$ probability for liability in civil cases like *Smith v Rapid Transit Inc.* that require liability be established by *the preponderance of the evidence*, and $\geq 90\%$ probability for guilt in criminal cases that require guilt *beyond reasonable doubt*.

Scholars have thus proceeded to investigate more complicated explanations of the reaction against the adequacy of the evidence in cases like *Smith v Rapid Transit Inc.* The standard move is to appeal to “cleaned up” toy cases that leave essentially

¹ Thomson (1986) motivates her pioneering discussion of legal evidence with this case and, more recently, Smith (2018) and Gardiner (forthcoming) also discuss this case.

² Smith (2018, footnote 3) lists a number of other legal cases resting on a single piece of apparently statistical evidence. Roth (2010, footnote 50, 51) notes cases that rest more or less exclusively on DNA evidence, referred to as “pure cold-hit DNA cases”. We will introduce one such case later.

³ Our definition here of “naked statistical evidence” is intentionally vague. Later we will argue that the properties of “naked” and “statistical” and the associated problems for legal proof are a matter of degree. The cases that best demonstrate the problems are those at the extreme end of the spectrum, but other cases near to the extreme serve just as well.

⁴ As revealed by a significant amount of empirical research since the early papers of Kahneman and Tversky (e.g., 1973).

no room for dispute about what is the appropriate probabilistic reasoning, and whether the relevant probability thresholds are met.⁵ Contra so-called *legal probabilism* (see Hedden & Colyvan, 2019), it is supposed that there is some property of evidence that is needed for legal proof, beyond high probability, that is lacking in cases of naked statistical evidence. The idea is that the key property can be isolated by finding pairs of toy cases for which the probability of guilt/liability is the same, but which elicit different reactions vis-à-vis the verdict. The challenge is then to articulate the difference between the evidence in the cases in question.⁶ In Sect. 1 we present some of the more prominent (pairs of) toy cases. But, unlike others, we do not take the reactions to these cases as our starting point for investigation.

While there is a place for investigating what we call *extra-probabilist* accounts of reasoning, by way of explaining reactions to legal cases like *Rapid Transit*, it is important that more straightforward probabilist accounts are not overlooked. After all, there are significant costs to any departure from the standard probabilist model of rational choice, whether in the legal or other public decision-making settings.⁷ That is reason enough to proceed cautiously in analysing provocative toy cases. Indeed, in Sect. 1 we argue that the specific toy cases that have been developed to investigate naked statistical evidence are potentially misleading. We thus—in Sect. 2—return to some of the real cases of naked statistical evidence that prompted the ensuing academic debate. Specifically, we provide what we take to be more faithful characterisations of these real cases. We go on to propose an explanation of the reaction to these cases that appeals only to traditional probabilist standards of reasoning in the legal context, albeit more nuanced aspects of this standard than is generally appreciated.

2 The problem with the popular toy cases

Let us introduce a couple of the well-known toy cases that purportedly provide clear-cut examples of “naked statistical evidence”.

Blue Bus. Mrs. Brown is run down by a bus on Victoria Street; 60 percent of the buses that travel along this street are owned by the blue bus company, and 40 percent by the red bus company. The only witness is Mrs. Brown, who is colour-blind. Given the lack of further information, one could argue that there is 0.6 probability that Mrs. Brown was run down by a blue bus and so the blue bus company should be held liable.

⁵ There are exceptions. Some scholars propose that the probabilistic reasoning and/or the threshold for meeting the burden of proof are not well appreciated in (real and/or toy) cases of naked statistical evidence, e.g., Schoeman (1987), Roth (2010), Di Bello (2019), Bolinger (2020).

⁶ For particular proposals to this effect or surveys of proposals, see, for instance, Thomson (1986), Redmayne (2008), Enoch et al. (2012), Pritchard (2015, 2018), Blome-Tillmann (2015), Gardiner (2019), Smith (2018).

⁷ E.g. one no longer has a single elegant principle for rational decision making, namely “maximise expected utility”, but rather specific principles for specific domains of decision making. Moreover, the specific principles may get complicated.

Prisoners. One hundred prisoners are exercising in the prison yard. Ninety-nine of them suddenly join in a planned attack on a prison guard; the hundredth prisoner plays no part. There is no evidence available to show who joined in and who did not. One prisoner is randomly selected from the yard. One could argue that there is 0.99 probability that this prisoner participated in the attack and so should be found guilty for this crime.⁸

In each of these cases the probabilities in question are—by construction—undoubtedly sufficiently high to meet the relevant standard of proof, insofar as such standards involve probability thresholds.⁹ And yet, as in the real cases on which they are modelled, there is resistance to award damages in the Blue Bus case and to convict in the Prisoners case. Something seems to be missing in the evidence. What is the missing ingredient?

These toy cases are duly designed to abstract away the many complications that plague real legal cases and so highlight the evidential reasoning at issue. It is not always clear, however, as to what are the mere complications that ought to be abstracted away. Moreover, it is not always clear whether and how readers embellish an abstract case.¹⁰ Blue Bus and Prisoners ostensibly rule out routine problems of probabilistic evidential reasoning. But in this respect, we claim, these toy cases may be rather misleading, for the reasons just given.

For example, in Blue Bus the probabilities about road usage are presented as both accurate and relevant to the question of the blue bus company's liability. The reader is not invited to challenge the frequencies cited, nor interrogate the reference class upon which they are based; whether, for instance, the reference class incorporates road usage at any time of day or is rather restricted to road usage at roughly the time of day of the incident. As such, the reader is directed to look for different potential deficiencies of the evidence, not its probative value for the liability of the blue bus company. That is a valuable project, but not necessarily the one most pertinent to a case like *Rapid Transit*. Moreover, we question whether the intuitions of readers about Blue Bus have been sufficiently examined for confounding influences that were not intended in the description of the case and have to do with the presentation of the evidence. After all, the reader is being asked to do something very unreasonable: we are being asked to accept the data about the relative frequencies of blue and red buses without question. Worse still, we are asked to make a decision about liability in this case without seeking further data—data that would be very easily

⁸ Both this Prisoners example and the Blue Bus example are presented in Redmayne (2008). Prisoners is attributed to Nesson (1979) while Blue Bus is inspired by Thomson (1986).

⁹ Note that one can always engineer the numbers, such that, for instance, in the Prisoners case, there are one thousand prisoners in the yard and 999 of them are involved in the attack. In this case the probability of guilt for a randomly-selected prisoner is 0.999.

¹⁰ For instance, interpretation issues are a point of debate regarding toy cases used to demonstrate probabilistic fallacies (see, e.g., the discussion of the “Linda problem” concerning the so-called conjunction fallacy in Tentori et al., 2004). The rejection of certain aspects of toy cases is perhaps most discussed in relation to the elicitation of moral judgments and is termed “imaginative resistance” in this context (see, e.g., Liao & Gendler, 2015).

obtained, without incurring any significant costs. For example, information about bus timetables for the route and the relevant time of the day would be very valuable here. Perhaps the blue buses do not run at the time the accident occurred or none of the blue bus routes encompass the site of the accident. It is at least possible that it is our reluctance to accept these stipulations that drives our wariness to award damages.

The case of Prisoners is similar. For starters, the body of evidence is itself quite remarkable: 99 out of 100 prisoners *certainly* participated in an attack, and yet nothing is known about the identity of the one prisoner not involved. The defendant was furthermore *randomly selected* from the group of 100 prisoners. Does that kind of setup mimic and clarify the evidential scenario in a case like *Rapid Transit*? We think not. And even if we were interested in Prisoners for its own sake, can it be expected that readers will arrive at judgments that duly take account of the stipulated features of the case? Even if the reader were to accept what is known about the prisoners in the yard, they may find it hard to accept that further evidence could not be easily sought, such as video evidence that might help to identify the one prisoner not involved in the attack. Moreover, in this case there is another complication arising from an ambiguity with the specification of the thought experiment. If only one prisoner is found guilty, an issue of justice arises. It would be very unfair to convict one prisoner when all one hundred have the same probability of guilt. Indeed, given the wording of the thought experiment, this seems to be a very natural reading. In this case, it is easy to confuse a reluctance to find guilty an individual prisoner because of suspicions about naked statistical evidence with a reluctance to find guilty because of equity issues. In order for the thought experiment to isolate naked statistical evidence, we need to ask whether it is reasonable to find all one-hundred prisoners guilty. Setting aside our earlier concerns about the availability of further evidence, there is arguably not such a strong intuition against finding all one hundred guilty. At least, we do not have this intuition, which suggests it cannot be taken for granted.¹¹

We do not here dispute the toy-case methodology. Or at least, we do not dispute this methodology understood in a “minimalist” sense, whereby cases simply serve to make a theoretical point vivid and are used to elicit everyday judgements (Machery, 2017).¹² The goal may be to determine the “common view” on the point at hand, which arguably has *some* normative, as well as social scientific, significance. But

¹¹ There is an interesting further question of the relationship between cases of naked statistical evidence, such as Prisoners, and the lottery paradox. To be sure, there are some similarities but there are also some differences (e.g., the lottery paradox is about what one ought to believe about a given ticket, whereas Prisoners is about a legal *decision*: the first is purely epistemic and the second is decision theoretic). It would take us too far afield to discuss such issues in full here but we see no reason to think that our treatment of cases involving naked statistical evidence should be easily adapted to an account of the lottery paradox. Our account, however, should be readily adaptable to any situation where “meta-uncertainty” might reasonably be thought to arise.

¹² Machery (2017) contrasts the “minimalist” method of cases with the “exceptionalist” and “particularist” methods which he claims not to be viable methods at all. On these latter methods, the judgements elicited in response to toy cases are treated as having a special status such that they bear directly on the formal or material facts under consideration.

even if there is good reason for extensive use of toy case methodology in the social and moral sciences, that does not mean that any given application of this methodology is beyond criticism. Just like empirical experiments that are intended to illuminate some phenomenon of interest, toy-case experiments may be compromised by confounders. Above, we questioned whether Blue Bus and Prisoners successfully distil the evidential conundrum of real cases like *Rapid Transit*. Moreover, we gave specific reasons to doubt that the reader takes on the Blue Bus and Prisoners toy cases as intended. We suggested that this may be because the details of the probabilistic calculations in the real cases are not, after all, “mere complications”.¹³ To the extent that such doubts are not addressed (say, by more extensive testing, involving comparisons of subtly different cases), any conclusion drawn about the nature and import of naked statistical evidence in the law on the basis of more and less widely shared judgements about these toy cases is at best tentative.

Our worry about drawing hasty conclusions regarding naked statistical evidence on the basis of the usual toy cases (when not sufficiently scrutinised) is compounded by a further move in the literature: the pairing of cases like those above with counterpart cases. For instance, Blue Bus is typically contrasted with the following case:

Blue Bus Testimony. Mrs. Brown is run down by a bus on Victoria Street; it is assumed that 50 percent of the buses that travel along this street are owned by the blue bus company, and 50 percent by the red bus company. The only witness is Mrs. Brown, who testifies that she was run down by a blue bus. Given the circumstances, it is estimated that Mrs. Brown has reasonably reliable vision: If the perpetrator was in fact the blue bus company, there is 0.6 probability that she would get this right, and 0.4 probability that she would get it wrong. (Likewise for Mrs. Brown’s reliability if the perpetrator was in fact the red bus company.) Given the lack of further information, one could argue that there is 0.6 probability that Mrs. Brown was run down by a blue bus and so the blue bus company should be held liable.¹⁴

Many report less discomfort with awarding damages in Blue Bus Testimony as compared to Blue Bus. (This is known as the “Wells effect” because it was purportedly first established by Gary L. Wells’ (1992) pioneering empirical studies.) Taken at face value, this result serves to further pinpoint the properties of “naked statistical evidence” and, accordingly, why it is deficient. Whatever this problematic evidence amounts to, Blue Bus has it while Blue Bus Testimony does not. Or so it is commonly accepted in the literature.

It is certainly plausible that the preliminary conclusion drawn from the Blue Bus pair and other similar pairs of cases—that there is some *qualitative* difference

¹³ We commented on ways that toy cases can be misinterpreted in footnote 10 above. Machery (2017) refers to the “explicit” versus the “implicit” content of a case, and notes that the latter can be difficult to regiment. See too Sorensen (1992) for discussion of the many reasons why the judgements elicited by toy cases may be confused or misleading.

¹⁴ This example appears informally in Redmayne (2008), following Thomson (1986) and with reference to cases developed by Wells (1992). The wording here is similar to that in Gardiner (2019).

between the evidence in the two cases that is important for legal proof—is correct. But the misgivings about real cases of naked statistical evidence, like *Rapid Transit*, may not be primarily due to any such qualitative difference. Moreover, given the boldness of the claim that qualitative aspects of evidence matter in ways that go beyond probative value, we suggest that there has been relatively little follow-up testing for potential confounders with respect to people’s judgements.¹⁵ The doubts we raised above about how people really understand and internalise the toy cases still stand. Moreover, the pairing of cases brings to light further worries. For instance, perhaps the presentation of the evidence in the two cases produces differing *priming* effects that have nothing to do with stable properties of the evidence. For instance, in Blue Bus Testimony, blue buses may simply be a lot more salient than red buses, whereas in the original Blue Bus, both types of buses are prominent in our reasoning.¹⁶ There may be further possible reasons for the asymmetric reaction to Blue Bus and Blue Bus Testimony that have nothing to do with the explicit, intended features of the two cases (such as privileging visual evidence).

Given this scope for doubt, it seems at best premature that the problem of explaining reactions to “naked statistical evidence” has more or less been rephrased in the literature as the problem of explaining the qualitative difference between the evidence in, e.g., Blue Bus and Blue Bus Testimony, that is pertinent to legal proof. This has become *the* accepted starting point for investigation.¹⁷ The proposal that we here put on the table is orthogonal to this line of investigation. We return to “square one” - we go back to real legal cases—in order to reassess what are paradigmatic cases of “naked statistical evidence” and what is the primary problem with such evidence. As will be seen, our proposal does not rest on any qualitative property of evidence; indeed, we will argue that the “naked statistical” nature of evidence comes in *degrees*, and provides indirect rather than direct reasons for doubting that the relevant legal standard of proof has been met.

3 The primary problem with “naked statistical evidence”

We have already briefly outlined one real case that provokes questions about “naked statistical evidence”: the case of *Smith v Rapid Transit Inc.* Our whittled-down description of the case above captures what we take to be the important evidential details; this description effectively serves as a representative toy case involving naked statistical evidence. This is a good case for our purposes, at least as we

¹⁵ A fairly recent study by Arkes et al. (2012) surveys the handful of empirical studies available at the time. As far as we are aware, there has not been much empirical work on the topic since.

¹⁶ We note that Arkes et al. appeal to priming effects to explain other non-standard aspects of our reasoning which they label “reasoning biases” as opposed to interesting challenges to legal probabilism. Admittedly, this is because their evidence suggests that people form correct subjective probabilities in both Blue Bus cases, yet incorrect subjective probabilities in these other cases. One might question, however, whether these experiments are successful in eliciting *subjective probabilities*; perhaps subjects simply appeal to the probabilities given in the case descriptions, whether or not these probabilities represent their degrees of belief in the relevant hypotheses.

¹⁷ Most of the papers cited earlier adopt this starting point.

described it, because the reasoning is relatively straightforward: one might assume that the only company running a bus route along a particular street will own most of the buses on that street, and is thus most probably (> 0.5) the company (if the offending bus is modelled as randomly selected) that was involved in an accident.

We seek at least one further real case to ground our inquiry. Consider this much-discussed candidate:

Sally Clark. When Sally Clark's two sons died suddenly (but separately) from no obvious natural cause, she was convicted of murder, despite there being no other incriminating evidence. A paediatrician testified at the trial that the probability that her two sons would have died of SIDS (unexplained natural causes) was 1 in 73 million. (That figure was criticised, with a revised estimate being in the vicinity of 1 in 9 million, still a very small number.) The jury found Sally Clark guilty in 1999, but this verdict was overruled in 2003, due in part to misgivings about the nature of the evidence in the original trial.¹⁸

The evidential reasoning in the Sally Clark case, however, involves too many complicating factors. There are many reasons to be unhappy with the statistics in this tragic case and these are well documented (e.g., see Dawid, 2005), and Balding & Steele, 2015). For a start, SIDS deaths are not well understood but it is known (and was known at the time) that such deaths are not independent, as assumed in the original calculation: one SIDS death in a family means that a second is more likely. It is also notable that the probability of someone like Sally Clarke committing double homicide was never calculated—it was simply assumed to be more likely than double SIDS deaths. In short, this is a case of *bad* statistical evidence and does not serve our purposes. What we need is a case of naked statistical evidence where there are no *obvious* flaws in the statistical reasoning—a case where it seems that the statistics is as good as it gets, so to speak, and yet there is apparently something missing.¹⁹

Next consider a case of DNA evidence:

Cold-Hit DNA. In 1972, a young nurse was raped and murdered in her San Francisco home. For various reasons, the case stalled for 30 years. In 2004, the case was reopened, and DNA found inside the victim was compared with 338,000 DNA profiles in California's offender database. The search yielded just one match, or "cold hit" to then-seventy-one-year-old John Puckett. The primary evidence against Puckett was the DNA match, although jurors also heard that he lived in the Bay Area in 1972, and had a 1977 sexual assault conviction. The government's DNA expert reported that the chance that a random

¹⁸ This case is discussed in Dawid (2005).

¹⁹ Another famous case found in the literature on this topic is that of the sentencing of the drug mule Shonubi (Colyvan et al., 2003; Colyvan & Regan, 2007; Tillers, 1997, 2005). But this case too has many complicating factors and is not the best case for isolating naked statistical evidence. See the references just noted for details of the complications we have in mind.

person from the population would match the profile was 1 in 1.1 million. The jurors found Puckett guilty, but many express reservations about this verdict.²⁰

This seems a good candidate for a case that can be straightforwardly characterised as an example of naked statistical evidence. One complicating factor is that, while the DNA-match evidence looks to be extremely *probative* (strong) evidence for Mr Puckett’s guilt, it is not sufficiently probative (even at face value) to clearly establish guilt beyond reasonable doubt. According to the DNA statistical model in question, the false positive rate is in the range of 10^{-6} . The true positive rate is assumed to be 1. The relevant likelihood ratio for the positive DNA match with Mr. P. is thus extremely large. In this case there is not merely a positive DNA match with Mr. P., but also a negative DNA match with all others in the database. (There was *just one* DNA profile in the database that matched that of the crime scene sample: Mr. P.’s.) This means the likelihood ratio for the total match evidence is even larger; the extent of increase depends on the probability that the database includes the culprit.²¹ This extremely large likelihood ratio for the total match evidence is to some extent counterbalanced, however, by what is plausibly an extremely low *prior* or base-rate probability for Mr Puckett’s guilt, assuming that the culprit could equally well have been any one of a large group of people, of which those in the database constitute only a small subset.²² All this is to say that the further evidence mentioned above is crucial for the case against Mr Puckett. We nonetheless treat this as an example of “naked statistical evidence” because the one statistic concerning the DNA match plays a considerable role in establishing high probability of guilt.

What is common to Rapid Transit and Cold-Hit DNA that makes a liability or guilty verdict, respectively, seem unwarranted? To begin with, it is worth noting that whatever is the issue with the evidence in these cases, it seems to be orthogonal to what sets Blue Bus apart from Blue Bus Testimony. Rapid Transit does closely resemble Blue Bus (in that it seems to rest on a “base-rate” or “general” statistic regarding cases of a certain kind). But Cold-Hit DNA, which we are grouping

²⁰ The description of this case closely follows that of Roth (2010), who discusses the rise of cold-hit DNA cases, involving larger and larger databases, in criminal law. Here we use this case simply as an example of one in which DNA evidence plays a significant role. The term “cold hit” in fact means something more specific: that the suspect was identified via a search process whereby a database of DNA profiles is checked for any matches with the crime scene DNA profile. This way of identifying a suspect, however, is not germane to our discussion. Note that DNA evidence does not strike everyone as problematic. Enoch (2012, page 221, footnote 38), for instance, refer to the “DNA exception to the usual suspicion with which statistical evidence is viewed”. But many nonetheless have reservations about guilty verdicts that rest largely on DNA evidence, and rightly so, as we will go on to argue.

²¹ For further details regarding the probative value of the total DNA match evidence in this sort of case, as well as the probative value of total DNA match evidence in other sorts of cases where the total DNA match evidence has somewhat different form, see Urbaniak and Di Bello (2021, esp. section 2.2), and references therein, especially Balding and Donnelly (1996).

²² Note that the *posterior* (or post-evidence) probability ratio for guilt versus innocence is, by probabilist reasoning, a product of the *prior* probability ratio for guilt versus innocence multiplied by the *likelihood ratio* associated with the evidence. That is (where G represents guilt, $\neg G$ represents innocence, and E represents the evidence that is learnt):

$$\frac{Pr_{\text{post}}(G)}{Pr_{\text{post}}(\neg G)} = \frac{Pr_{\text{prior}}(G)}{Pr_{\text{prior}}(\neg G)} \times \frac{Pr_{\text{prior}}(E|G)}{Pr_{\text{prior}}(E|\neg G)} = \frac{Pr_{\text{prior}}(G|E)}{Pr_{\text{prior}}(\neg G|E)}$$

together with Rapid Transit, looks a lot like Blue Bus Testimony (in that it seems to rest on a “trace” or “individualised” statistic arising from the particular case in question).²³ We treat both these real cases—or rather, both these “toy” characterisations of real cases—as exemplars of “naked statistical evidence”.

The defining feature of “naked statistical evidence”, we propose, is that it amounts to a single line of evidence, albeit one that seems to offer, on the basis of a salient statistical model, a very precise and relatively high probability of guilt or liability. The problem is that there can be legitimate uncertainty about the statistical model; hence the misgivings about finding a defendant guilty or liable on the basis of such evidence. Recall that this sort of uncertainty is precisely what the well-known toy cases rule out by stipulation. Hence our concern that these particular toy cases have foreclosed important lines of investigation regarding naked statistical evidence. To be clear, we are not saying that whenever some evidence is described in terms of a single statistic, as in the well-known toy cases, it necessarily constitutes “naked statistical evidence” that is vulnerable to uncertainty about the reliability of the model in question. The well-known toy cases—Blue Bus and Prisoners—cite what a reader may reasonably take to be reliable statistical evidence. Perhaps there was no opportunity for human error in gathering the relevant data, or the cited statistic is robust across multiple possible models.²⁴ Rather our point (refer back to Sect. 2) is that, to the extent that the toy cases rule out the possibility of error in the statistical model, they may not adequately account for the misgivings that people have about *real* legal cases that rest largely on statistical evidence. Moreover, we argued that there is reason to doubt that readers accept the intended reading of the toy cases—that the cited statistics are watertight—not because they could not possibly be so but because statistics are perceived, perhaps not entirely consciously, to be typically vulnerable to error.

To elaborate: In cases of naked statistical evidence, the probative value of the evidence with respect to guilt or liability or else the probability of guilt or liability given the evidence, is ostensibly expressed in the statistical statement in question. For example, in the cold-hit DNA case above, the probability of a false positive is in the range of 10^{-6} , according to the DNA analysis in question. (This false positive rate plays a significant role in the likelihood ratio for the total match evidence.) The fact that the probability of a false positive is not zero tells us that there is some, but not considerable, uncertainty about guilt here. But there is another kind of uncertainty that we might reasonably ask after and which falls outside the scope of this statistical model: we might be uncertain about the assumptions on which the statistical model is based.

What we have in mind here is a kind of meta-uncertainty. Arguably, such uncertainty is not resolved by making the probabilities delivered by the statistical model smaller or larger, as the case may be. Typically we need to construct another statistical model, or else we must simply appeal to brute expert judgement, to quantify

²³ We by no means wish to suggest that the meaning of “base rate/general” versus “trace/individualised” evidence is settled by our remarks here. This is the subject of ongoing discussion, which, as mentioned, we sidestep in this paper.

²⁴ We discuss the relative robustness of an evidence-based inference in more depth in Sect. 4.

this meta-uncertainty (concerning, e.g., mistakes in the DNA laboratory or contamination of the crime scene).²⁵ In extreme cases, model uncertainty can completely invalidate the model. A model for predicting the population abundance of a species in a given region may deliver a value, with the uncertainty about the value also specified by the model. But if we find that the model is built on the wrong causal structure, say, then all bets are off about the results of the model. Suppose that in our cold-hit DNA case we found that the algorithm employed in the DNA match analysis was at odds with our best DNA science. Here we do not tend to revise our probability of a false positive from 10^{-6} to, say, 10^{-5} . Rather, we recognise that the model that delivered 10^{-6} is flawed and we no longer have confidence in any results this model delivers.

There can also be uncertainty about the connection between the statistical model and the guilt/liability hypothesis in question. That is, the assumptions regarding the *relevance* of the statistical model may be in doubt. For instance, there may be uncertainty about the appropriate *reference class* for calculating probabilities pertinent to a trial (Colyvan et al., 2003). In our Rapid Transit case, for instance, one might question whether it is reasonable to treat this accident as involving any one of the bus companies, randomly selected, that may have been using the street on a typical day. That is, one might question the reference class grounding the estimate that Rapid Transit was most probably the company that caused the accident. While this was the only company that officially operated there, perhaps at different times of the day the prevalence of other bus companies on the street varied, and was sometimes quite high. Or perhaps the day in question was an unusual one whereby many other bus companies were in operation throughout the city. Moreover, perhaps bus companies diverge significantly with respect to accident rates and this should have had some bearing on the reference class and the probabilities for liability derived from it.

The issue is that naked statistical evidence is typically derived from a particular statistical model and, as such, is susceptible to undermining meta-uncertainty. Without some assurances that there is no meta-uncertainty or, alternatively, that the meta-uncertainty in question has been appropriately dealt with, the probabilities delivered by the statistical model are best-case scenarios: the probability of a false positive is 10^{-6} , provided there were no errors in the lab, no crime-scene contamination, the statistical analysis took the appropriate form, and so on.

²⁵ Such meta-uncertainty is well known in model-based science, one instance of which is known as *model uncertainty*. This is uncertainty about the model itself and is a kind of systematic uncertainty (related to systematic error). Model uncertainty can be notoriously difficult to quantify (Regan et al., 2002) and may be difficult to reconcile with standard Bayesian epistemology depending on how it is understood (Weisberg, 2015).

4 Meta-uncertainty: digging deeper

We have given a basic account of the worries raised by naked statistical evidence. But one might regard this a rather incomplete explanation. We have not yet said why naked statistical evidence, *in particular*, raises issues of meta-uncertainty. Indeed, our account of meta-uncertainty above might suggest it undermines any inference from a body of evidence to a verdict. That is, our appeal to meta-uncertainty might be thought to prove too much: it seems that not only verdicts based on naked statistical evidence are open to doubt, but any verdict based on any evidence can be similarly undermined. It behoves us to offer a more detailed account of meta-uncertainty that reveals why it is of particular concern in the case of naked statistical evidence.

We start by characterising meta-uncertainty in a way that shows it does in fact affect *any* line of evidence. That is, it is not an on/off issue, triggered by properties that only some evidence possesses. Nonetheless we go on to meet the challenge of explaining what sets naked statistical evidence apart. Both the attributes of this evidence heighten problems of meta-uncertainty: the fact that the evidence is “statistical” or based on a salient model, and the fact that the evidence is “naked” or lacking “redundancy”. These features will be discussed in turn.

Our characterisation of meta-uncertainty draws on the work of Kadane and Schum (1996). They suggest that the probative value of any given evidence with respect to some hypothesis of interest depends on oft-suppressed intermediary links in the “inferential chain” concerning the credibility of the evidence and its relevance to the hypothesis. These intermediary links can be interpreted as meta-uncertainty about the *prima facie* bearing of the evidence on the hypothesis. In the case of naked statistical evidence, the bearing of the evidence on the hypothesis is mediated by a statistical model, and the meta-uncertainty concerns the aptness of that model.

Our Cold-Hit DNA case will serve to illustrate. The inferential chain is plausibly as per Fig. 1.

On a naive treatment of the DNA-match evidence, the intermediary links concerning credibility and relevance are ignored. Indeed, reasoning about DNA-match evidence typically does not go beyond such a naive treatment. According to our story above, for instance: “The government’s DNA expert reported that the chance that a random person from the population would match the profile was 1 in 1.1 million.” This is to assume that the DNA sample in question was indeed that of the perpetrator, and that it was moreover analysed in an error-free way. But what if the crime scene was contaminated with DNA that was not that of the perpetrator, and errors were introduced in the recording and testing of this sample DNA? This is the meta-uncertainty that Fig. 1 makes explicit.

On the naive model the evidential impact of E (for Cold-Hit DNA) is measured as follows²⁶:

²⁶ To keep things simple in our discussion in this section, we appeal to the likelihood ratio associated with an isolated match between the suspect’s DNA and the crime scene DNA. We will ignore the further evidence that there was just one positive DNA match—the suspect’s—amongst the samples in the given database (refer to footnote 21).

$$\frac{\Pr(E|H)}{\Pr(E|\neg H)} = \frac{\Pr(R|H)}{\Pr(R|\neg H)} = 1/0.00000091 \approx 1,100,000.$$

How is this modified once we pull apart E and R to account for the meta-uncertainty associated with the inferences from E to C and/or from C to R ? We can reason about the meta-uncertainty more explicitly by expanding the left-hand expression above:

$$\begin{aligned} \frac{\Pr(E|H)}{\Pr(E|\neg H)} &= \frac{\Pr(E|R\&H)\Pr(R|H) + \Pr(E|\neg R\&H)\Pr(\neg R|H)}{\Pr(E|R\&\neg H)\Pr(R|\neg H) + \Pr(E|\neg R\&\neg H)\Pr(\neg R|\neg H)} \\ &= \frac{\Pr(E|R\&H)}{0.00000091.\Pr(E|R\&\neg H) + 0.99999901.\Pr(E|\neg R\&\neg H)}. \end{aligned}$$

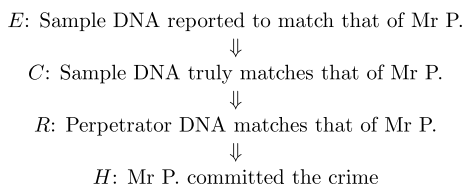
Before we assumed that $\Pr(E|R\&H)$ —the value of the numerator, since H being true entails that R is true—was equal to one. But now we recognise that even if the perpetrator’s DNA matches that of Mr P., this need not mean that there is a recorded match between the DNA of Mr P. and the sample DNA from the crime scene. Perhaps the sample DNA taken from the crime scene was not that of the perpetrator, or perhaps, even if it was, errors were made in its analysis. So the numerator may be much less than one. Moreover, the denominator may be much greater than 0.00000091. Before we assumed that $\Pr(E|R\&\neg H)$ was equal to one and $\Pr(E|\neg R\&\neg H)$ was equal to zero. But the latter may be significantly greater than zero, since even if there was no actual match between the perpetrator’s DNA and that of Mr P., there may well be a recorded match between the sample DNA and that of Mr P., again due to either poor sampling or poor analysis of the sample. In all, the likelihood ratio in question may be a great deal smaller than 1,100,000.

Note that it may not always make sense to think of meta-uncertainty as affecting the likelihood ratio associated with some evidence. Consider the Rapid Transit case. We can again think of the meta-uncertainty as hidden links in the inferential chain, as per Fig. 2. The first link concerns the credibility of the evidence and the second link concerns its relevance for the hypothesis in question. Here again, there is a tendency to conflate E with R .

But in this case it is somewhat unnatural to model the import of the evidence in terms of how it changes a prior probability distribution, as per the likelihood ratio. That is because here E arguably determines the very base-rate or prior probability for H . One must account for the meta-uncertainty or hidden inferential links in the prior probability. The crucial question in the Rapid Transit case concerns the aptness of the (roughly specified) bus prevalence statistic, and how well it reflects the probability that any given bus company (such as Rapid Transit) was responsible for the accident involving Mrs. Smith. Answering this question involves further statistical modelling and assumptions (e.g., concerning whether or not there is variation in bus prevalence at a finer grain, say for different times of the day, and whether or not any given bus is equally likely to be involved in an accident). Such assumptions lie in the background of the above line of reasoning and can be made explicit and, of course, brought into question.

Finer details aside, this characterisation of meta-uncertainty is consistent with probabilist or Bayesian reasoning, and applies to the evidence in Cold-Hit

Fig. 1 Cold-hit DNA inferential chain



DNA, Rapid Transit, and to any other kind of evidence. Indeed, while the term “meta-uncertainty” may sound like a matter of higher-order or secondary judgment about first-order probabilities, we model it in a way that bears directly on first-order probabilities. That is intentional and it sets our account—with its emphasis on intermediary links in the inferential chain—apart from other proposals that raise merely secondary doubts about the probability estimates involved in cases of naked statistical evidence. For instance, Moss (2018) argues that the probabilities derived from naked statistical evidence, high as they may be, fail to constitute the “probabilistic knowledge” required for legal proof. Others claim that these probabilities, again high as they may be, fail to meet a further criterion of “diversity of evidence” (e.g., Cohen, 1977), “completeness of the evidence” (e.g., Kaye, 1986), or “stability of belief” (e.g., in a manuscript by Günther that appeals to a model owing to Leitgeb, 2017). While these proposals are in the same ballpark as ours, the further judgments they appeal to—whether a version of meta-uncertainty, diversity, completeness, or stability—are directed at, rather than informing, first-order probabilities. As such these proposals are more controversial because they invoke dubious epistemic properties and amount to a greater departure from ordinary Bayesian reasoning.

Why then, is this a good account of the qualms people have about verdicts based on naked statistical evidence, *in particular*? The lesson thus far seems to be that we should second-guess all evidential reasoning, since there is always the possibility of further links in the inferential chain, representing meta-uncertainty, that may undermine the calculated probative value of the evidence.

As noted above, we *do* think there is reason to second-guess the evidential reasoning supporting any legal verdict. Nonetheless verdicts based on naked statistical evidence are peculiar, or rather, they are “extreme” cases on the spectrum. The first reason is the “statistical” nature of this evidence. It is not that we think “statistical” is a well-defined property of evidence; merely that it tends to be used to describe evidence that is associated with a well-known reference class or some other salient model from which the relevant probability is derived. Our proposal is simply that, in these cases, meta-uncertainty — which typically involves murkier reasoning about possibilities that are difficult to quantify — can be overshadowed by the more straightforward model-based inferences. For instance, in the face of sophisticated population data regarding the prevalence of certain patterns of DNA in the population, it is easy to overlook, say, the possibility of crude human error in recording the DNA information in a database. By contrast, where evidence is not associated with a well recognised model, considerations of meta-uncertainty arguably do not stand apart as less salient and are thus not so readily overlooked. For instance, even the most simple-minded

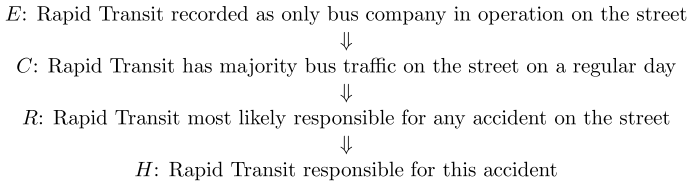


Fig. 2 Rapid transit inferential chain

interrogator would presumably not treat testimonial evidence as fact, but rather consider the ability and will of the witness to report events accurately.²⁷

Perhaps the bigger issue, when it comes to meta-uncertainty, is the “naked” aspect of verdicts based on naked statistical evidence. Recall that we take “naked” here to refer to there being just a *single* line of evidence supporting the guilt hypothesis in question. Roughly speaking, verdicts based on a single line of evidence are more vulnerable to challenge. Multiple lines of evidence can build in redundancy and thus serve to “back-up” a verdict in the face of challenges from meta-uncertainty. The kind of back-up provided depends not just on the prima facie probative value of each line of evidence taken in isolation, but also on the extent to which the lines of evidence are interdependent, and whether the meta-uncertainty or source of doubt in question targets what is common to, or rather what differentiates, the lines of evidence.

We will not attempt a fully general analysis of the interactions between multiple lines of evidence. We will simply appeal to an example that illustrates the kinds of principles at play. The example is described in Fig. 3. Here we have three lines of evidence originating, respectively, in raw evidence $E1$, $E2$, and $E3$. Assume, as per Fig. 1, that H concerns the guilt of Mr. P. The first chain involving $E1$ is, let’s say, the one described in Fig. 1.

We consider the extreme case where each of $E1$, $E2$, and $E3$ is treated as equivalent to R . That is, there is, at least initially, no concern about the credibility or relevance of this raw data with respect to H . Let us further assume that, taken in isolation, the prima facie likelihood ratio for each of $E1$, $E2$, and $E3$ is 10.²⁸ Given that it is ultimately R that matters for the truth of H , and all the pieces of raw data are assumed to be equivalent to R , there is redundancy in this body of evidence: just one of the pieces of evidence would suffice to deliver the same posterior probability for H . It is this redundancy that serves as insurance against meta-uncertainty worries.

The precise insurance provided by multiple lines of evidence is admittedly somewhat complicated. Let us focus on the first line of evidence involving $E1$, and consider what insurance is provided by $E2$ or $E3$. It all depends on the nature of the meta-uncertainty or the possible error that comes to light. In our simple case, there are three

²⁷ Having said that, some contend, on the basis of empirical data, that eye-witness testimony is unreliable and this unreliability is not properly appreciated by the courts or the general public (Arkowitz & Lilienfeld, 2010; Buckhout, 1980).

²⁸ In this case, if the prior probability for H were 0.5, its posterior, based on one (or all) of the pieces of raw evidence, would be quite high: a little larger than 0.9.

possibilities. (i) It may be the first link in the inferential chain that is doubtful, from $E1$ to C_A : the inference from the purported DNA sample match to there being an actual DNA sample match. Perhaps there was a laboratory error. In this case, either $E2$ or $E3$ serves as back-up. From Fig. 3, we see that neither of these lines of evidence shares the link between $E1$ and C_A . We can interpret $E2$ as a different laboratory report of the DNA match. Just because one laboratory analysis may have been corrupted does not mean that the second laboratory analysis was similarly corrupted. (ii) It may be the second link in the inferential chain that is in doubt, from C_A to R . Perhaps a dubious sample was taken from the crime scene. In this case, $E2$ provides no insurance, since it concerns the same sample (the link from C_A to R is shared). But $E3$ provides insurance, as this line of evidence involves a different sample, and its inferential chain does not involve the link from C_A to R . (iii) It may be the third link between R and H that is doubtful. Perhaps DNA match analyses are based on a faulty scientific model, such that even when the implementation of the analysis is error-free, it is not clear what can be inferred from a positive match. In this case, neither $E2$ nor $E3$ provides insurance, since all lines of evidence involve this link.

Our example is stylised. One can think of variations—ways in which a line of evidence may provide some, but not complete back-up for another line of evidence that is threatened by doubts about credibility or relevance. The details can get complicated but the stylised example suffices for our purposes. One can see that multiple lines of evidence allow for a certain amount of redundancy or back-up in establishing a high probability for some hypothesis (here the guilt of a suspect) being true. This high probability is then resilient to certain meta-uncertainty challenges. Referring back to our example: even if the credibility of $E1$ were radically undermined (represented by the link from $E1$ to C_A), if all else remains unchanged (and given the assumptions stated above) then the probability of the guilt of Mr P. remains unchanged. Another way to put it is that multiple lines of evidence mean that a verdict is less *sensitive* to the precise meta-uncertainty profile. Only some meta-uncertainty profiles, or possibilities for error, will undermine the verdict; others will not, due to the back-up. Verdicts based on naked or single-line evidence lack this resilience or insensitivity to error. Moreover, our example suggests that it may be difficult to remedy this problem when one lacks a diversity of *sources* of evidence. Figure 3 depicts a strategy for making naked statistical evidence less naked, but there are limits to this strategy. It is difficult to effectively back-up all the inferential links associated with a single source of evidence, like a DNA match.

In sum, verdicts based on naked statistical evidence are particularly vulnerable to meta-uncertainty challenges because (a) the nature of the evidence is such that meta-uncertainty tends to be under-estimated (so that even checks of the sort illustrated in Fig. 3 may not have been conducted), and (b) the verdict is highly sensitive to any changes in the meta-uncertainty profile. But naked statistical evidence is not alone in this respect.

It is also worth noting that there are other ways to ensure one's statistical findings are robust (i.e., not overly sensitive), in the sense of not being too dependent on

Fig. 3 Multiple reinforcing inferential chains

$$\begin{aligned} E1 &\Rightarrow C_A \Rightarrow R \Rightarrow H \\ E2 &\Rightarrow C_A \Rightarrow R \Rightarrow H \\ E3 &\Rightarrow C_B \Rightarrow R \Rightarrow H \end{aligned}$$

the specifics of a particular statistical model. For instance, one might eschew singular statistical models in favour of suites of statistical models, where each model in the suite is consistent with known data but differs on other details (such as using different reference classes or different assessments of the reliability of the eye-witness testimony). Standard sensitivity analysis counsels us to trust only results shown to hold across all the relevant models. Or, we might be more permissive and trust results that hold across most of the models in the suite. Alternatively, we might aggregate the models in question, perhaps by considering the weighted average of their results.²⁹ These are all ways of shoring up, at least to some extent, the findings of statistical models. Such approaches do, however, face serious problems. For instance, a great deal depends on which models are thought to be relevant and thus included in the suite. If we are too liberal and admit any model consistent with known data, we will find that we have very few, if any, robust results; too restrictive and almost every result is deemed to be robust. Moreover, we need a principled way to decide which models are deemed relevant and which are not. If we are going to use weighted averages of the results of the models, the results will be hostage to the specific weights used so, in each case, the weights will require rigorous and independent justification. Such problems are, perhaps, not insurmountable, but discussing such issues further would take us too far afield. For present purposes, we are content to note that there are other approaches in the spirit of our proposal but we take our proposal to be less of a departure from orthodox Bayesian methodology.³⁰

5 Insufficient probability of guilt after all?

So what exactly, on our proposal, is the normative lesson highlighted by misgivings about verdicts based on “naked statistical evidence”? Is it quite simply that guilty/liable verdicts in these cases are wrong because, due to meta-uncertainty, the probability of guilt/liability does not surpass the relevant threshold for legal proof? It may seem that if one wants to uphold legal probabilism—the probability-threshold interpretation of legal standards of proof—then this is the only lesson that can be drawn. But it amounts to a rather bold claim: that in all cases of naked statistical evidence, the probability of guilt/liability is, despite appearances, too low. Call this the *Insufficient Probability* claim.

Our conceptual points about meta-uncertainty do not, on their own, establish Insufficient Probability. At the very least, however, there is a more modest version of this lesson that is consistent with legal probabilism: the idea is simply that verdicts based on naked statistical evidence should trigger alarm bells because they *might* be undermined by meta-uncertainty. In such cases one should proceed cautiously and

²⁹ Indeed, one might argue that naked statistical evidence is best thought of in terms of weighted averages of permissible statistical models, in which case naked statistical evidence has a certain robustness built in from the get go (similar to the multiple lines of evidence described in Fig. 3). We thank an anonymous referee for this interesting suggestion.

³⁰ For Bayesians, this is reason enough to explore the approach we take; for others, our approach is perhaps, at best, an interesting exercise in Bayesian epistemology.

check all assumptions that underpin one's inferences about guilt/liability. To give an analogy: when I realise that I am booked on the very last flight that would get me to my destination in time for an extremely important meeting, I double-, even triple-check my flight time and re-confirm my lift to the airport. It may turn out that my doubts were unfounded and the chain of transportation, in this case, as opposed to the chain of inference, was already in order. But I should recognise the vulnerability of my plans to different types of hitches and ensure that I have properly taken the risks into account.

We go further in claiming that the risks associated with naked statistical evidence will generally be such that its probative value, on a reasonable assessment, is much less than appearances suggest. Hence, this evidence will generally raise the issue of Insufficient Probability after all. Take the case of Mr P. While some potential sources of meta-uncertainty may be easily resolved, such as human error in the analysis of a DNA sample (a backup analysis is viable, and may have already been performed, as per *E1* and *E2* in Fig. 3 above), other sources of meta-uncertainty may not be so easily resolved. For instance, it may be difficult to rule out the possibility that the sample DNA was *not* that of the culprit, perhaps because it is unclear whether or not there were past errors in sample storage. Assume in the case of Mr P. that it is as likely as not that the sample DNA was *not* that of the culprit. Taking this meta-uncertainty into account would then reduce the likelihood ratio concerning Mr P.'s guilt, with respect to the DNA match evidence, by at least one half.³¹ That is a considerable drop in the probative value of the DNA match evidence. It would mean that the final probability of Mr P.'s guilt may well be insufficient for a guilty verdict.

It might be asked how one can accurately estimate meta-uncertainties, such as the DNA sampling error in the case of Mr P. Is our running estimate that it is as likely as not that the sample DNA is that of the perpetrator a reasonable one? It is not obvious. This may seem an exaggerated probability of error. But note that even a lower error rate would still significantly reduce the likelihood ratio associated with the evidence of a reported match between the sample DNA and that of Mr P. For instance, if there was just a 25% chance of storage error, the probative value of the match evidence would still be considerably lower (the associated likelihood ratio would be approximately 25% of its *prima facie* value). Moreover, we have so far been assuming any errors to be *innocently* made. If one rather suspected *deliberate* manipulation of the data, then reasonable estimates of error might be a lot higher.

These remarks raise an interesting issue regarding legal probabilism. The sort of meta-uncertainty that plagues real cases of naked statistical evidence (and indeed, much of the uncertainty that arises in legal cases more generally) is hard to estimate. In such circumstances, there will not be an obvious *uniquely correct* probability estimate, given the evidence at hand, for the guilt/liability of the defendant. For one

³¹ The numerator of the expanded likelihood expression is $\Pr(E|R\&H)$. Before we assumed it equalled one, since we assumed that if the perpetrator's DNA matched that of Mr P, then it would indeed be reported that the sample DNA matched Mr P. But now we are suggesting that it is as likely as not that the sample DNA is not that of the perpetrator, so the probability of a reported match between the sample DNA and that of Mr P drops to approximately 0.5.

thing, it is not generally accepted that there is always a uniquely correct probability for a hypothesis given some evidence.³² Even if there were, however, the probability would in many cases be a matter of persistent disagreement. Only in rare cases, such as those involving random selection from a population, such as in the Prisoners toy case, is there a salient uniquely correct probability. (This is another inadequacy of the popular toy cases—the special nature of the evidence obscures important details regarding the probability-threshold interpretation of legal standards of proof.) So legal probabilism—the probability-threshold interpretation of legal standards of proof—will not be straightforward to implement.

By way of offering a clearer picture of this issue of implementation, we appeal to the notion of a *reasonable* probability judgement. This is a probability judgement that may be adopted by a careful and conscientious inquirer. With the exception of some special cases involving random selection, there would not be a uniquely reasonable probability for a hypothesis given some evidence; rather there would generally be a range of reasonable probabilities.³³ That is vague, but it may be the most one can say in the abstract. Plausibly, what counts as a reasonable probability judgement must be contested on a case-by-case basis. One might expect more agreement about what is the range of reasonable probability judgements than about what is the uniquely correct probability judgement in any given case. The problem of implementation arises for legal probabilism when the range of reasonable probabilities for guilt/liability straddles the probability threshold for the relevant standard of proof.³⁴

Plausibly, contested cases of naked statistical evidence raise precisely this implementation problem for legal probabilism. In the case of Mr P., we saw that reasonable, albeit fairly pessimistic, estimates for sample-storage error rates reduce the likelihood ratio considerably (such that, plausibly, the inferred probability of Mr P.'s guilt would not surpass the very high threshold for criminal conviction). Nevertheless, more optimistic estimates for error rates also seem reasonable and may yield a probability for Mr P.'s guilt that *does* surpass the very high threshold. Consider too the Rapid Transit case. Here we suggest that concerns of relevance—about whether the accident is properly seen as one of many accidents that could equally be caused by any bus on the road—could reasonably mean the probability of liability does not meet the threshold of 0.5, despite there also being reasonable estimates for this probability that do surpass this threshold.

So what is the appropriate implementation of legal probabilism—the appropriate legal verdict—in the problematic threshold-straddling circumstances? We cannot hope to here offer a full response to this very general interpretative question for legal probabilism. We will instead simply suggest some alternatives in relation to naked statistical evidence. The first is the least radical position as regards the interpretation

³² This is a point of debate amongst contemporary epistemologists.

³³ The range of reasonable probability functions would presumably amount to a convex set, but we will not explore such details here.

³⁴ The other possibilities for the range of reasonable probabilities are i) both the lower and upper bounds of the range fall below the probability threshold and ii) both the lower and upper bounds fall above the probability threshold. Legal probabilism straightforwardly requires an innocent verdict in the former circumstances and a guilty verdict in the latter.

of legal probabilism. It is simply to acknowledge disagreement. The range of reasonable probability estimates simply represents the extent of reasonable disagreement about a defendant's guilt. Where this range straddles the probability threshold for the relevant burden of proof, different verdicts can be justified. In the case of Mr P., a guilty verdict could be justified on the basis of legal probabilism, let's say, but so too could an innocent verdict. The controversy raised by verdicts of guilt/liability based on naked statistical evidence would then be seen to mark persistent reasonable disagreement about what is the right verdict. The normative lesson is simply that these cases are not cut and dried and therein lies the deficiency of naked statistical evidence: Some reasonably find the evidence sufficiently probative for an unfavourable verdict; others reasonably do not.

The above may be unsatisfying for those who think that legal standards of proof should yield a single verdict that accommodates what is, after all, *reasonable*, disagreement. Assuming we are right that such disagreement can be represented by a set of probability functions over the relevant hypotheses, this would require a greater departure from the standard interpretation of burdens of proof under legal probabilism. The 'threshold test' would not be a simple matter of whether some precise probability estimate for guilt/liability is greater than the threshold probability. Rather, the test would concern a range or set of reasonable probabilities. There are two salient ways such a threshold test might go, in terms of what is necessary and sufficient for a guilty/liable verdict: (i) when *at least one of* the reasonable probability judgements surpasses the relevant threshold, and (ii) when *all of* the reasonable probability judgements surpass the relevant threshold.³⁵ The second clearly amounts to a more stringent standard of proof; it would plausibly not be met in most cases of naked statistical evidence due to reasonable meta-uncertainty doubts. This interpretation thus fits better with misgivings about naked statistical evidence. It is not that *no* reasonable probability, given the evidence, is sufficiently high for guilt/liability. It is simply that *there exist* reasonable probabilities of guilt/liability that are not sufficiently high. And the lesson would be that in these circumstances, a guilty verdict is not justified.

Let us take stock, then, of the normative lesson we propose to take away from the widespread misgivings about guilty/liable verdicts based on naked statistical evidence. Our claim is that the probative value of this evidence can be reasonably contested, since it is difficult to effectively remove the meta-uncertainty, or find "back-ups" for all the crucial links in the evidence's inferential chain. At the very least further checks on the meta-uncertainty and hence the probative value of the evidence are in order. In most cases, we suggest, the remaining reasonable disagreement will be such that on at least some reasonable assessments, the probability of guilt/liability is below the relevant threshold. Hence guilty/liable verdicts based on naked statistical evidence will generally not be uniquely justified. We leave open how legal probabilism should be spelled out in cases where the set of probability estimates for the probability of guilt/liability—our representation of reasonable disagreement—"straddles" the relevant threshold. Either an innocent/non-liable verdict

³⁵ I.e. subvalue and supervalue, respectively.

is uniquely justified, or else both this verdict and the opposing verdict can be justified. That is a broader issue that goes beyond the scope of this discussion.

6 Concluding remarks

In closing, we return briefly to how our diagnosis of the problem of “naked statistical evidence” compares with other proposals in the literature. As mentioned, most authors seek a qualitative difference between statistical and other kinds of evidence, and see pairs of toy cases like Blue Bus and Blue Bus Testimony as successfully isolating this difference. Some authors argue that the difference concerns some epistemic property (whether “causal connection” (e.g., Thomson, 1986, Colyvan & Regan, 2007), “sensitivity” (e.g., Enoch et al., 2012), “safety” (e.g., Pritchard, 2018), or the like) that is important for legal proof, over and above high probability. Others argue that the difference concerns the moral costs of inferential error—that the costs surprisingly depend on the *type* of supporting evidence (e.g., Bolinger, 2020). These proposals regarding the problem of statistical evidence are orthogonal to the one we propose in this paper (i.e., our proposal is not inconsistent with there being *further* problems of statistical evidence along any of these lines). We claim just that these proposals do not capture the most basic or primary lesson to be learnt from misgivings about real cases of naked statistical evidence in the law. The primary lesson is one of refining our ordinary probabilistic reasoning in the legal setting to take proper account of meta-uncertainty.

Our meta-uncertainty diagnosis does not suppose that naked statistical evidence is a distinctive form of evidence. This evidence rather lies at one end of a continuum, with respect to the extent of mismatch between *prima facie* probative power and redundancy or “back up” in the evidence. Naked statistical evidence is striking because it typically suggests a high probability of guilt or liability (as the case may be) despite being a single line of evidence. It is moreover puzzling because it is based on a model that is perceived to be objective. But whenever there is a single line of evidence—e.g., testimony from a single eye-witness—similar issues of meta-uncertainty arise and one should have similar reservations about the reliability of the evidence. Our account is thus partially revisionary, since some scholars, at least, see misgivings about naked statistical evidence as indication that it is a special class of evidence that requires special consideration in the law.

Despite our account of naked statistical evidence being somewhat revisionary, it is worth noting that our proposal sits well with important aspects of standard legal practice. The law usually asks after *means*, *motive*, and *opportunity*. On the account we’re advancing here, such requests can play an important role in shoring up naked statistical evidence. Providing plausible accounts of means, motive, and opportunity need not be thought of as extra requirements taking priority over the statistical evidence; rather, they’re basic checks that things are in order with the statistical model.³⁶ More generally, we can look for ways to “triangulate” on a legal verdict.

³⁶ Other informal checks might take the form of plausible causal stories underwriting the statistical findings (Colyvan et al., 2003).

We can't always guard against mistakes in the laboratory or contamination of the crime scene but we can check that the accused had the opportunity to commit the crime in question. This, in turn, gives us indirect evidence that the statistical model is reliable.

Acknowledgements Thanks to David Balding, Christopher Birch, Renee Bolinger, Mark Burgman, Nevin Climenhega, Scott Ferson, Georgi Gardiner, Susan Haack, Brian Hedden, Anne Ruth Mackor, Deborah Mayo, David Neil, Mike Redmayne, and Andrew Robinson for helpful discussions on the topic of this paper. Earlier versions of this paper were presented to The School of Politics, Economics and Philosophy at the University of York (May 2012), The Department of Mathematics and Statistics at the University of Melbourne (April 2015), The Department of Philosophy at the University of Sydney (May 2015), The Australian Society of Legal Philosophy annual conference at the University of Sydney School of Law (June 2015), the Australasian Association of Philosophy annual conference at the Macquarie University (July 2015), the Center for Philosophy of Science at the University of Pittsburgh (September 2015), The Department of Philosophy at the University of Miami (November 2016), the Munich Centre for Mathematical Philosophy at the Ludwig-Maximilians University, Munich (January 2018), the workshop "Evidence in Statistical, Biomedical and Forensic Sciences" at the Polytechnic University of Marche, Faculty of Medicine, Ancona, Italy (June 2018), the Department of Philosophy at the Australian National University (October 2018), the Department of Philosophy at Stanford University (November 2018), the Department of Philosophy at Utrecht University (June 2019), the 2019 Munich-Sydney-Turin Philosophy of Science Conference Statistical Reasoning and Scientific Error at the Carl Friedrich von Siemens Foundation, Munich (July 2019), the Department of Philosophy at the University of Wollongong (September 2019), the Princeton-Rutgers Foundations of Probability Seminar Series, Rutgers University (November 2020), and the Institute for Risk and Uncertainty at the University of Liverpool (June 2021). We'd like to thank the audiences in each of the forums for their many helpful comments, suggestions, and criticisms. Mark Colyvan's research on this paper was funded by an Australian Research Council Discovery grant (grant number: 180103549) and a Carl Friedrich von Siemens Research Award of the Alexander von Humboldt Foundation. Katie Steele's research on this paper was funded by an Australian National University Futures Grant and an Australian Research Council Discovery grant (grant number: 170101394).

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Arkes, H. R., Shoots-Reinhard, B., & Mayes, R. S. (2012). Disjunction between probability and verdict in juror decision making. *Journal of Behavioral Decision Making*, 25(3), 276–294.
- Arkowitz, H., & Lilienfeld, S. O. (2010). Do the eyes have it? *Scientific American Mind*, 20(7), 68–69.
- Balding, D. J., & Donnelly, P. (1996). Evaluating DNA profile evidence when the suspect is identified through a database search. *Journal of Forensic Sciences*, 41(4), 603–607.
- Balding, D. L., & Steele, C. D. (2015). *Weight-of-evidence for forensic DNA profiles* (2nd Edn.). Chichester, UK: Wiley.
- Blome-Tillmann, M. (2015). Sensitivity, causality, and statistical evidence in courts of law. *Thought*, 4(2), 102–112.

- Bolinger, R. J. (2020). The rational impermissibility of accepting (some) racial generalisations. *Synthese*, 197(6), 2415–2431.
- Buckhout, R. (1980). Nearly 2,000 witnesses can be wrong. *Bulletin of the Psychonomic Society*, 16(4), 307–310.
- Cohen, L. J. (1977). *The Probable and the Provable*. Oxford: Clarendon Press.
- Colyvan, M., & Regan, H. M. (2007). Legal decisions and the reference class problem? *International Journal of Evidence and Proof*, 4(1–2), 274–285.
- Colyvan, M., Regan, H. M., & Ferson, S. (2003). 2001. ‘Is it a Crime to Belong to a Reference Class?’, *The Journal of Political Philosophy*, 9(2), 168–181. Reprinted. In H. E. Kyburg & M. Thalos (Eds.), *Probability is the Very Guide of Life* (pp. 331–347). Chicago: Open Court.
- Dawid, A. P. (2005). Statistics on trial. *Significance*, 2(1), 6–8.
- Di Bello, M. (2019). Trial by statistics: is a high probability of guilt enough to convict? *Mind*, 128(512), 1045–1084.
- Enoch, D., Spectre, L., & Fisher, T. (2012). Statistical evidence, sensitivity, and the legal value of knowledge. *Philosophy and Public Affairs*, 40(3), 197–224.
- Gardiner, G. (2018). Evidentialism and moral encroachment. In K. McCain (Ed.), *Believing in Accordance with the Evidence: New Essays on Evidentialism* (pp. 169–195). New York: Springer Verlag.
- Gardiner, G. (2019). Legal burdens of proof and statistical evidence. In D. Coady & J. Chase (Eds.), *The Routledge Handbook of Applied Epistemology* (pp. 179–195). New York: Routledge.
- Gardiner, G. (forthcoming). Legal evidence and knowledge. In Larsson-Aarnio, M., & Littlejohn, C. (Eds.), *The routledge handbook of the philosophy of evidence*.
- Günther, M. to appear. ‘Probability of Guilt’.
- Hedden, B., & Colyvan, M. (2019). Legal probabilism: a qualified defence. *Journal of Political Philosophy*, 27(4), 448–468.
- Kadane, J. B., & Schum, D. A. (1996). *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. New York: John Wiley and Sons Inc.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.
- Kaye, D. H. (1986). Do we need a calculus of weight to understand proof beyond a reasonable doubt? *Boston University Law Review*, 66(4), 657–672.
- Leitgeb, H. (2017). *The Stability of Belief*. Oxford: Oxford University Press.
- Liao, S., & Gendler, T. S. (2015). The problem of imaginative resistance. In J. Gibson & N. Carroll (Eds.), *The Routledge Companion to the Philosophy of Literature* (pp. 405–418). New York: Routledge.
- Machery, E. (2017). *Philosophy Within Its Proper Bounds*. Oxford: Oxford University Press.
- Moss, S. (2018). *Probabilistic Knowledge*. Oxford: Oxford University Press.
- Nesson, C. R. (1979). Reasonable doubt and permissive inferences: the value of complexity. *Harvard Law Review*, 92(6), 1187–1225.
- Pritchard, D. (2015). Risk. *Metaphilosophy*, 46(3), 436–461.
- Pritchard, D. (2018). Legal risk, legal evidence and the arithmetic of criminal justice. *Jurisprudence*, 9(1), 108–119.
- Regan, H. M., Colyvan, M., & Burgman, M. A. (2002). A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications*, 12(2), 618–628.
- 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.
- Schoeman, F. (1987). Statistical vs. direct evidence. *Noûs*, 21(2), 179–198.
- Smith, M. (2018). When does evidence suffice for conviction? *Mind*, 127(508), 1193–1218.
- Sorensen, R. (1992). *Thought Experiments*. Oxford: Oxford University Press.
- Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, 28(3), 467–477.
- Thomson, J. J. (1986). Liability and individualized evidence. In W. A. Parent (Ed.), *Rights, Restitution and Risk: Essays in Moral Theory* (pp. 225–250). Harvard MA: Harvard University Press.
- Tillers, P. (1997). Introduction: three contributions to three important problems in evidence scholarship. *Cardozo Law Review*, 18(6), 1875–1889.
- Tillers, P. (2005). If wishes were horses: discursive comments on attempts to prevent individuals from being unfairly burdened by their reference classes. *Law, Probability, and Risk*, 4(1–2), 33–49.

- Urbaniak, R., & Di Bello, M. (2021). Legal probabilism. In Zalta, E. N. (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). <https://plato.stanford.edu/archives/fall2021/entries/legal-probabilism/>.
- Weisberg, J. (2015). Updating, undermining, and independence. *British Journal for the Philosophy of Science*, *66*(1), 121–159.
- Wells, G. L. (1992). Naked statistical evidence of liability: is subjective probability enough? *Journal of Personality and Social Psychology*, *62*(5), 739–752.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.