On the Epistemic Significance of Noise

Zoë Johnson King (Harvard University) and Boris Babic (University of Toronto and University of Hong Kong)[1]

> *Winston had nothing to do with the Lottery, which was managed by the Ministry of Plenty, but he was aware (indeed everyone in the party was aware) that the prizes were largely imaginary. Only small sums were actually paid out, the winners of the big prizes being nonexistent persons.*
>
> From George Orwell's *1984*

1. Introduction

Many epistemologists and philosophers of science have discussed so-called *statistical inference*, i.e. inference from an observation about an individual's category membership and evidence about the proportion of members of the category that bear a certain property to a conclusion about the probability that the individual herself possesses that property. Some have argued that statistical inferences face a host of purely epistemic problems, while others have argued that there are problems with these inferences that go beyond the epistemic — usually ethical and political problems of various kinds. Likewise, there is a longstanding legal literature on courtroom factfinding based on statistical evidence, which again divides into criticisms of uses of statistical evidence that are purely epistemic and those that are moral or political.

In this paper we join the epistemic camp. What is novel about this paper, however, is that we identify and explore an epistemic problem about an aspect of statistical inference that has been widely taken for granted by authors across the board. It is widely taken for granted in the literature that statistical evidence can easily be *highly probabilifying* (i.e. that it can support an arbitrarily high credence in a proposition about a target event — we will explain this shortly), with efforts to locate an epistemic or ethical flaw concentrated elsewhere. Indeed, it is sometimes thought that to deny that statistical evidence can be highly probabilifying would defy Bayesian orthodoxy. However, in this paper we argue that realistic cases in which statistical evidence is highly probabilifying are in fact surprisingly rare. And we argue for this conclusion using standard Bayesian theoretical machinery, thus taking precisely the approach thought to be bound to give statistical evidence the green light.

This means that, although our account provides no reason to reject any extant epistemological or ethical accounts of problems with statistical inference, it does suggest that in a great many cases it is not necessary to advert to them in order to explain what is wrong with such inferences. A simpler diagnosis is available: one need not reach for additional theoretical bells and whistles to describe a problem with a certain inference if that inference's premises do not support its conclusion in the first place. And, we suggest, it is in fact surprisingly difficult to find real-world cases in which the fact that X is F and the observation that $a/100$ of observed Fs have been found to be G jointly support a conclusion that X is a% likely to be G or a credence of $0.a$ in the proposition that X is G. We observe that the way in which problems of statistical inference are often framed in the legal and philosophical literature involves a surreptitious slide between two quite different types of inference, and we argue that, once the problem is correctly framed, in most realistic examples it is no longer the case that the available

[1] Both authors contributed equally to the writing of this paper at all stages of development.

statistical evidence supports a high credence in the proposition in question. So, although it is true *in theory* that statistical evidence can be highly probabilifying, in practice things are rarely as neat as the literature suggests. This significantly limits the applicability of conclusions based on simplified theoretical assumptions.

2. Setup

Consider the following three cases:

> RAFFLE.[2] You decide to participate in a raffle at your local school fête. There are 100 tickets. You buy 10 of them. The raffle is drawn and you go, excitedly, to hear the result.
>
> PRISON YARD.[3] 100 prisoners are in a yard wearing identical outfits (with hoods and masks) so that they cannot be visually distinguished from one another. 90 of the prisoners team up and attack the guards. The remaining 10 prisoners try to defend the guards.
>
> IPHONE THEFT.[4] You leave the seminar room to get a drink, and you come back to find that your iPhone has been stolen. There were only two people in the room, Jake and Barbara. You have no evidence about who stole the phone, and you don't know either party very well, but you know (let's say) that men are 10 times more likely to steal iPhones than women.

Now, here are three propositions:

> P1. You have not won the raffle.
> P2. Prisoner Number 37, who was randomly selected from among the 100 prisoners, participated in the attack.
> P3. Jake stole your iPhone.

Based on the evidence in RAFFLE, PRISON YARD, and IPHONE THEFT, what should be your credence in each of P1, P2, and P3?

At first blush, the cases might look structurally similar. After all, the winning ticket must be one of the 100 tickets in the raffle, only 10 of which you hold, and it was randomly selected from among them. Similarly the 90 assailants are somewhere among the 100 prisoners who were in the yard, from which group Prisoner Number 37 was randomly selected. On this evidence, P1's probability is 0.9 and P2's the same. Likewise the thief must be someone who was in the seminar room while you were gone, making Jake and Barbara the only options. And their respective genders, coupled with the statistical information that you somehow know, might appear to suggest that Jake is 10 times as likely to be the thief as Barbara is – thus P3's probability on your evidence appears to be 10/11 (i.e. 0.91). These cases appear similar because they all seem to

[2] This is a version of the much-discussed "lottery paradox". See Kyburg (1961) for the original version, and see Hawthorne (2004) for a book-length treatment of the problem.
[3] This case is originally from Nesson (1979) and has been widely discussed ever since. It is structurally similar to the Gatecrasher case in Cohen (1977), which is also widely discussed. (N.B. we have changed the numbers to preserve mathematical similarity across our three cases. Also, in Nesson's original case the one prisoner who does not attack the guard instead goes and hides behind a shed. We have altered this odd detail to make it clear that the innocent prisoners are good guys.)
[4] This case is lifted word-for-word from Buchak (2014).

involve inferences from a fact about an individual's membership in a reference-class and an observation about the proportion of members of the reference-class that bear a certain property to a conclusion about the probability that the individual possesses the property. Indeed, PRISON YARD and iPHONE THEFT appear to involve almost identical reasoning, as follows: 90% of prisoners in the yard participated in the attack, and Prisoner Number 37 was in the yard, so Prisoner Number 37 is 90% likely to have participated in the attack; 91% of iPhone thefts are committed by men, and Jake is the only suspect who is a man, so the thief of this particular iPhone is 91% likely to be Jake.

Many scholars have found these apparent structural analogies illuminating. In particular, there is a longstanding tradition in legal scholarship of expressing strong reservations about findings of guilt or liability against people in positions like that of Prisoner Number 37 in PRISON YARD or Jake in IPHONE THEFT, and many philosophers have thought that the apparent structural similarity of both cases to RAFFLE can explain these reservations. After all, it is often held that the statistical evidence in a case like RAFFLE licenses high credence in P1 but not outright belief. Moreover, if you do believe P1 on the basis of the statistical evidence in RAFFLE, then intuitively you do not *know* P1 even if it turns out to be true. Indeed, a belief in P1 on the basis of the statistical evidence in RAFFLE would violate standard sensitivity and safety conditions on knowledge, since the winning ticket could easily have been one of your tickets and if it were then you would still have possessed the statistical evidence and would still have believed P1 on its basis. Nor would your belief have any causal connection to the fact that P1 is true, if indeed it is a fact, were your belief to be based on the statistical evidence alone. And any of these traditional epistemological problems might also be said to arise in PRISON YARD and IPHONE THEFT; an outright belief in P2 and P3 based on the statistical evidence in these cases seems inappropriate, seems not to constitute knowledge even if the propositions are true, is both insensitive and unsafe, and is causally unrelated to the facts. Philosophers have accordingly taken various of these traditional epistemological problems to explain the aversion to verdicts based on statistical evidence found in legal scholarship, even granting *arguendo* that the statistical evidence in PRISON YARD and IPHONE THEFT would indeed license credences of 0.9 and 0.91 respectively, in light of these cases' apparent similarity to RAFFLE (see e.g. Thomson 1986, Redmayne 2008, Enoch, Spectre & Fisher 2012, Buchak 2014, Moss 2018, Pardo 2018, Pritchard 2018, Gardiner forthcoming).

Indeed, for some of these philosophers the assumption that statistical evidence licenses high credence in cases like PRISON YARD and IPHONE THEFT is crucial for the argumentative work that they want such cases to do. For example, Buchak (2014), from whom we have taken the IPHONE THEFT example word-for-word, explicitly compares this case to a version of RAFFLE and argues that the statistical evidence licenses high credence but not outright belief in both (pp.292-293). Buchak's approach distinguishes the kind of evidence that licenses high credence from the kind that licenses full belief, arguing that legal verdicts and other attributions of blame are only appropriate based on outright beliefs (rather than credences) on the grounds that statistical evidence licenses high credence but seems intuitively insufficient for blame or punishment (pp.299-301). Buchak takes these considerations to support *belief-credence dualism*: the thesis that full beliefs are not reducible to credences. Similarly, Moss (2018) uses the intuitive insufficiency of statistical evidence as a basis for blame and punishment to support her view that credences can constitute knowledge. Moss's argument is based on the same intuitive data: that the statistical evidence in PRISON YARD seems insufficient for conviction and likewise that in Buchak's IPHONE THEFT case the statistical evidence seems to be an insufficient basis for blaming Jake (pp.206-207). Moss's explanation likewise grants that the statistical evidence in these cases is highly probabilifying – indeed, Moss is helpfully explicit about this, stating that "statistical evidence can justify a factfinder in having an arbitrarily high credence that a defendant is guilty" (p.205). Her account follows Buchak's up to this point, deviating instead by offering a different explanans for the explananda: Moss cuts belief out of the picture and argues that the high credences licensed by statistical

evidence in PRISON YARD and IPHONE THEFT themselves fail to constitute knowledge, and that, since knowledge of probabilistic contents is the standard for legal proof, they would be an inappropriate basis for legal verdict (pp.208-213).

Other philosophers and legal scholars have simply taken for granted that statistical evidence licenses high credence, looking elsewhere for explanations of our aversion to legal verdicts – or similar normative beliefs and negative attitudes formed outside the courtroom – based on such evidence. The literature here is enormous. For a very modest sample, see Tribe (1971) and Nesson (1985) for well-known arguments eschewing probabilistic calculation in the legal process, Cohen (1977) for an alternative to traditional probability theory whose calculations supposedly deliver intuitively acceptable verdicts in cases involving statistical evidence, and Brook (1982), Wasserman (1991), Colyvan, Regan and Ferson (2002), and Pundik (2008) for moral and political concerns about statistical inferences that do not focus on these inferences' epistemic credentials.

In this paper we want to emphasize that cases like IPHONE THEFT are in fact far *less* closely analogous to cases like RAFFLE and PRISON YARD than they are widely taken to be. Importantly, we do not deny the structural parallel between RAFFLE and PRISON YARD (or the prison yard case's companion case – that of the gatecrashers).[5] On the contrary, we want to emphasize that a great deal of artificial precision has been introduced to PRISON YARD in order to preserve its structural similarity to RAFFLE. In both cases we are confronted with a set – the raffle tickets in one case, the prisoners in the other – in which a precise, known proportion of members must possess a certain property – being the winning ticket in one case, having participated in the attack in the other. There is zero chance that this property is possessed by a non-set-member, given the stipulations of the cases. And there is zero chance of mistake as to the precise proportion of members of the set that do in fact possess the property; again, it is part of the stipulations of the cases that exactly one ticket will win and exactly 90 prisoners participated in the attack. Furthermore, the set members are themselves perfectly *interchangeable* – which is to say that, based on the available evidence, no single one of them is any more or less likely to possess the property of interest than any other. And the inference that we must draw concerns a member of this very set, as opposed to a non-member whose possession of the property of interest might be predicted based on past observations.

Things are already very different from this simple picture when we turn to IPHONE THEFT. Here the only set such that a precise proportion of its members must possess a certain property is the set of students who were in the seminar room when you went to get a drink, one of whom must be the one who stole your iPhone. (Indeed, even this is not quite true, since Jake and Barbara might be colluding.) But applying the sort of reasoning that one finds in RAFFLE and PRISON YARD to *this* set would yield only a 0.5 probability that Jake stole the iPhone. What does all the work in allegedly rationalizing a 0.91 credence in Jake's guilt is an additional piece of evidence to which no analogue appears in RAFFLE and PRISON YARD: the statistic that men are 10 times as likely to steal iPhones as women. A better way to understand the statistical inference in IPHONE THEFT would therefore be to construe the set of interest as the set of *iPhone-stealers* and the property of interest as the property of *being a man*, since the statistic concerns the prevalence of this property among members of that set. But now much of the artificial precision is lost. For it is no longer part of the setup of the case that the property of interest cannot be possessed by non-set-members; iPhone-stealing men, iPhone-stealing women, non-iPhone-stealing men and non-iPhone-stealing women can all be found out there in the wild. Nor is the precise proportion of set members who

[5] We should emphasize: this means that we *do* think that high credence is licensed in cases like that of the prison yard and the gatecrasher. In such artificially precise cases, one must look beyond the present paper for a reason not to believe the defendant guilty, blame them, and/or convict them. A great many putative reasons can be found in the literature, some of which we have cited in the main text. Our aim in the present paper is not to find fault with any of the arguments in this literature but to observe just how unrealistic are the artificially-precise cases on which they are based, and how differently real-world examples work in practice.

possess the property built into the setup of the case; it is true that *if* the statistic stipulated in the setup is correct then we should expect the ratio of men to women among iPhone thieves to approach 10:1 in the aggregate, but this clearly does not mean that any randomly-selected group of 11 iPhone thieves must contain exactly ten men and one woman.

More to the point, in this case we should be much more cautious than in cases like RAFFLE and PRISON YARD about blithely assuming that the statistic stipulated in the setup is correct. The inference about Jake and Barbara is not a direct inference from the proportion of the interchangeable members of a set possessing a property to the probability that a randomly-selected member of this set possesses the property. It is instead a completely different type of inference: a predictive inference from past observations of previous iPhone thefts and the genders of their perpetrators to the probability that a newly-encountered iPhone thief is a man.[6] And a real-life version of IPHONE THEFT would raise a lot of pressing questions about the statistic that Buchak quixotically says you "know". Unlike in RAFFLE and PRISON YARD, it is not at all clear where this statistic came from. Is it God's word? A written report? The result of informal observations? A federal count of the last 10 years of reported iPhone thefts? The bold assertion of a podcaster during a rant about how men and women are psychologically different? Furthermore, even if we knew the source of the statistic, there would be additional questions to be resolved before we could appropriately model the evidence. Importantly, the case concerns the probability that someone *steals an iPhone*, but any real-world statistic about past thefts and their perpetrators can at most reflect the probability that someone *is caught stealing an iPhone* (since thefts that are not caught are not recorded, except perhaps by God). This means that, unlike in the simpler cases of direct inference, there is now a non-zero chance of mistake as to the proportion of set members possessing the relevant property. Some men might have been erroneously "found" to have stolen an iPhone (and might therefore be counted as thieves when in fact they are not), while some might have stolen iPhones and gotten away with it (and might therefore not be counted as thieves when in fact they are). And likewise for women. As a result, in order to reason well with this statistic we would need to know more about the data collection process. As a simple illustration: if the data come from police prosecutions, but we discover that the police only prosecute male thieves, then we can expect false negatives to be heavily lopsided and the statistic to be highly unrepresentative.

In short, as soon as we move away from clean-cut philosophical cases of statistical inference and into the real world, we introduce the *risk of misclassification*. This is the focus of the present paper. As we will discuss in the next section, even a fairly minor risk of misclassification can have a major impact on the resulting predictive inference. And this weakening of the weight of the evidence becomes progressively more drastic as the risk of misclassification increases.

This result means, we think, that philosophers should be cautious in applying theoretical results about artificially clean-cut cases to the real-life cases that they are supposed to illuminate. Stipulating away all risk of misclassification is an idealization that enables theorists to generate interesting results about fictional cases. But those results may not hold – or may look very different, at least – when we attempt to apply them in the real world.

3. Main Point

In this section we will model cases like PRISON YARD and IPHONE THEFT more carefully.

Consider again PRISON YARD, wherein we have 100 prisoners, 90 of whom participated in the attack. It is also assumed

[6] The terms 'direct inference' and 'predictive inference' are widely used in the philosophy of statistics. See Kyburg (1974) and Levi (1977) for influential early papers on direct inference.

that the prisoners are indistinguishable from each other – that is to say, we know that 90 participated in the attack, but there is no individuating information that would make any prisoner selected at random more or less likely to have been involved in the attack than any other. The question is: for any given prisoner, what is the probability that they participated in the attack?

In this case, we would say that we have a sample of $n = 100$ prisoners, and $x = 90$ of them participated in the attack. For each prisoner, this is an instance of a *Bernoulli process*: a data generating distribution (or 'likelihood', as Bayesians say) such that each observation falls into one of two categories (i.e. participated in the attack, did not participate in the attack). This process is governed by one parameter: the proportion of those who participated in the attack, which we will denote by θ. Hence, this is a common type of problem, known as a case of *univariate inference for a proportion*. As Bayesians, we are interested in θ, and our goal is to use our evidence – i.e., observations about x – in order to formulate credences about θ.

Given this setup so far, the likelihood is given by

$$\ell(x|\theta) \propto \theta^{x}(1-\theta)^{n-x}. \tag{1}$$

Where x corresponds to the sum of prisoners participating in the attack. But in this case we do not need to estimate θ, since it is stipulated as part of the setup that we know the true proportion of prisoners who participated in the attack. We know by hypothesis, that is, that $\theta = 0.9$. We can think of this as the *true population proportion*, since the only population of interest in this case is the prisoners who were in the yard and we are assuming that we know precisely what proportion of them participated in the attack. It is not as if the prison yard is a sample and we are using data from it to draw inferences about other prison yards or different places where people attack other people. On the contrary, in this case we are interested in only this incident and in only this prison yard. Hence, the probability that Prisoner Number 37 (or any other randomly-selected prisoner) participated in the attack simply reduces to $\Pr(X_{37} = 1) = .9^{1}(1 - .9)^{0} = 0.9$. This step is licensed by the fact that the prisoners are exchangeable, which is to say that the joint probability is unaffected by permutations of order. In other words, the same would be true for prisoner Number 23, prisoner Number 67, et cetera, since it is part of the setup of the case that the prisoners are indistinguishable from one another.

Notice that this is a profoundly uninteresting case of inference. We know the true population proportion and all of the prisoners are exactly alike. So, the question we are really asking in a case like PRISON YARD is like this: We have an urn with 100 marbles, 90 of which are blue and 10 of which are white. If you select a marble at random, then what is the probability that it is blue? The question is so simple that one might wonder what the trick is. But there is no trick. If we take the example as given, then that is really all there is to it; the probability of drawing a blue marble out of an urn that contains 100 marbles, 90 of which are blue, is 0.9. But real life is more interesting than drawing marbles from urns in known proportions, as we will soon see.

Visually, the data-generating distribution in PRISON YARD can be depicted as follows.

Figure 1: A Visual Representation of PRISON YARD.

This looks simplistic (because it is). But we start here because we will complicate the picture in the examples to come.

The first step in making this case more realistic would be to assume that while our evidence is unambiguous and unproblematic, it comes from a sample that is smaller than the full population. Our task would then be to make a prediction about a new member of the population that was not included in the sample. So, for example, suppose that in the last two prison yard guard attacks, 90 out of 100 prisoners participated in the attack. Now we observe a new, third prison yard attack on the guards. And we are interested in the probability that in this new fight, Prisoner Number 37 participated in the attack. The likelihood remains the same as in Equation (1), but it is no longer the case that we know for sure that θ = 0.9. Instead, we have prior evidence suggesting that θ is somewhere in the vicinity of 0.9. But that is far from clear, of course, since our evidence is based on the last two guard attacks in the prison yard rather than the present case. Visually, the situation is the same as in Figure 1, but θ is now unknown rather than being stipulated to be equal to 0.9.

Figure 2: A Visual Representation of PRISON YARD when θ is not known.

We have included Figure 2 to make the reasoning as transparent as possible to the reader. The first case was a case of a known proportion. This is a case of an unknown proportion. And soon we will see a case of an unknown proportion with noisy evidence. What is interesting about this case (Figure 2) is that with θ unknown we are more squarely in the world of

Bayesian inference, since we now need to identify a prior distribution for θ. In other words, we have to specify how likely we think all the possible values for θ are, before seeing the evidence (i.e. the two guard attacks), and we then have to update that prior distribution on the observed data from the two guard attacks using Bayes' Rule.

A common prior distribution for inference involving a proportion is called a *beta distribution*. It is given by

$$\pi(\theta|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (2)$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta) / \Gamma(\alpha + \beta)$ and $\Gamma(n) = n - 1!$.

The most important bit to focus on here is the component of the distribution that is a function of θ, namely $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. The term in front is what we call a *normalizing constant*: it is not a function of θ, but rather a multiplier that guarantees that the resulting distribution integrates to 1 – i.e., that it is a valid probability function. When we apply Bayes' rule, we effectively have to multiply $\theta^{\alpha-1}(1-\theta)^{\beta-1}$ with $\theta^{x}(1-\theta)^{n-x}$ and normalize the result. In such a case, the normalizing constant would become $B(\alpha + x, \beta + n - x)$ – indeed, this is the sum of the binomial coefficient of our likelihood and the beta term from the prior – and the posterior would be

$$\pi(\theta|\alpha, \beta) = \frac{1}{B(\alpha + x, \beta + n - x)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}. \quad (3)$$

Notice that this is likewise a distribution with a beta form, except that it now has updated parameters. As Johnson King and Babic (2019) and Babic et al (2021) emphasize, this equation lends itself to a natural interpretation: the parameters of the prior distribution, α and β, function like 'pseudo-observations' – i.e. imaginary observations – of prisoners, where $\alpha / (\alpha + \beta)$ corresponds to the 'pseudo-proportion' – i.e. the proportion of prisoners that we think it makes sense to assume would participate in an attack on the guards before seeing any evidence.

In order to draw inferences about what happens in the third attack on the guards, we have to make some assumptions about $\alpha / (\alpha + \beta)$. It is natural to suppose that $\alpha = \beta = 1$, and hence that $\alpha / (\alpha + \beta) = 0.5$. This corresponds to the uniform distribution for θ, which many scholars argue is reasonable in the absence of prior information. Suppose, then, that we assume that $\alpha = \beta = 1$ and subsequently observe two prison fights, in each of which 90/100 prisoners participate in the attack. Then our posterior distribution will be a beta distribution with new parameters $(1 + 90 + 90, 1 + 10 + 10) = (181, 21)$.

How about the probability that in the third guard attack, prisoner Number 37 participated in the attack? This is given by the predictive probability, which in the univariate case involving a proportion corresponds to the posterior mean, i.e. $(\alpha + x) / (\alpha + \beta + n)$. In our case, this is 181/202 = 0.89. Here we have applied Laplace's well-known Rule of Succession, which is really just a special case of predictive Bernoulli inference – namely, predictive inference for a Bernoulli process under a uniform prior for the proportion.

One might wonder why the prediction about prisoner Number 37 has dropped from 0.9 to 0.89. But this is in fact precisely what one should expect, since we started with a prior distribution that very weakly assumes *ex ante* that half of all prisoners in yards during guard attacks participate in the attacks, and we then observed two attacks in which most of the prisoners (90 percent) participated. As a result, we update our prior and move strongly toward 0.9, but we do not

quite reach 0.9 because the initial uniform prior is still holding us back a little bit.[7]

This situation is still very different from IPHONE THEFT. In IPHONE THEFT, we have prior evidence of the proportion of iPhone thefts committed by a man, presumably based on some sample that was previously obtained. So far this is like the situation in Figure (2), i.e. the modified version of PRISON YARD. But – and this is what most interests the authors – in most real-life cases the evidence is imperfect. By 'imperfect evidence', we mean that there is a possibility of some past misclassification. It stands to reason that at least some men and some women were falsely convicted of stealing iPhones, while other thieves were never caught. And, as noted in the previous section, in this case we are interested in the proportion of *actual* iPhone thefts perpetrated by men, but the most we can observe is the proportion of *recorded* thefts so perpetrated. (In criminology, this is the distinction between *offense rates* and *arrest rates*.) For the purposes of this project, then, when we use the terms 'noise' or 'misclassification', what we have in mind are classification errors in a Bernoulli process.

Misclassification can occur for very mundane reasons, such as a clerical error. But, as the above discussion suggests, we can also use it to model more substantive sources of data noise. For example, if a racially biased police force only enforces criminal laws in predominantly black neighborhoods, avoiding predominantly white neighborhoods, then we would expect to see arrest rates in black neighborhoods that are exaggerated relative to the true offense rates due to misclassification caused by the police force's enforcement bias.

IPHONE THEFT is therefore a case that involves two important deviations from PRISON YARD and RAFFLE. First, it is a predictive inference about a new observation that is outside of the original sample which constitutes our evidence (as is captured by the difference between Figure 1 and Figure 2). Second, the original sample that constitutes our evidence is itself probably imperfect – i.e., it probably contains a non-zero rate of misclassification (as will be captured by the difference between Figure 2 and Figure 3, which is below). This makes IPHONE THEFT more realistic, and more interesting, than PRISON YARD and RAFFLE. But it also means that in order to appropriately model IPHONE THEFT we cannot use the same strategy that we used for the simpler cases. We must further refine the tools that we developed for the modified version of PRISON YARD. To do this, we draw on a methodology first developed in Babic et al (2021) and more recently applied in Babic and Johnson King (2023).

[7] As we say below, we do not endorse the uniform prior, nor does our argument depend on it being the right prior in cases like this. We simply use it for illustration due to its long history and its relationship to Laplace's Rule of Succession.

Consider again the setup that we had in Figure (2). And let us suppose for simplicity's sake that we only have one kind of mistake: some men were falsely convicted of stealing an iPhone when in fact they did not. There is nothing special about this assumption, and indeed we will relax it later. We only make it for now in order to introduce as few additional parameters into our model at a time as possible, for ease of exposition.

To account for misclassification, then, we can introduce a parameter that captures the false positive error rate for men, as follows:

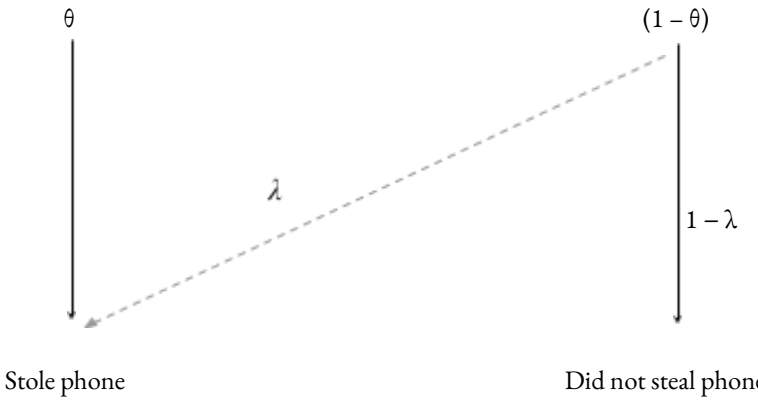

Figure 3: IPHONE THEFT with noise level λ.

Now we have quite a different situation. The evidence remains the same as that originally described by Buchak. But if we want to consider a realistic version of this kind of evidence then we must take seriously the fact that in reality it is inevitably imperfect; our model should not ignore this fact. This means that we can still proceed as before and think about what our credence that Jake stole the iPhone should be, but it is far less easy than drawing marbles from urns. Since we do not know who was falsely convicted, when we observe the 'data' that 91/100 iPhone thefts were perpetrated by men we should not rule out the possibility that any given one of these observations is mistaken, even if the probability that we assign to each mistake is very low.

We could consider the sum of all of these error probabilities – an approach that is developed in Winkler and Gaba (1990), Gaba and Winkler (1992) and Gaba (1993). But a more compact way to solve the problem would be to have a prior distribution for the false positive rate – as in Babic et al (2021). In Figure (3), the false positive rate is given by λ. We still have a Bernoulli process, but instead of phone-stealing men being drawn with rate θ, they are now drawn with rate $\theta + (1-\theta)\lambda$. That is, we have all the true phone-stealing men, plus the innocent men who were falsely convicted of stealing phones. Likewise for non-phone-stealers: instead of being drawn with rate $1-\theta$, they are now drawn with rate $(1-\theta)(1-\lambda)$ — the true non-phone-stealers plus those who stole and got away with it. Hence, our new likelihood is given by

$$\ell(x|\theta,\lambda) = [\theta + (1-\theta)\lambda]^{x}\,[(1-\theta)(1-\lambda)]^{n-x} \tag{4}$$

If we let $\varphi = \theta + (1 - \theta)\lambda$ then we could say that for each observation the process is Bernoulli in φ. And the likelihood reduces to $\varphi^{x}(1 - \varphi)^{n-x}$.

Now suppose that we observe 100 phone thefts, 91 of which are classified as being perpetrated by a man. We first need to specify a prior for θ. As we did in PRISON YARD, let's say that this prior is uniform. That is,

$$\pi(\theta) \propto 1. \tag{5}$$

It is not obvious that this is the best choice of prior. Although many scholars defend uniformity as a way to model indifference – since it seems like the most noncommittal prior – it is nonetheless a little strange to assume that every proportion is equally likely. A bell curve prior might be more plausible, for instance. And below we will consider a few different prior specifications. But, for simplicity's sake, we will proceed for now with a uniform prior.

However, we also need to specify a prior for λ. To do this we must ask ourselves: what is a reasonable false positive rate to assume *ex ante*? We could answer this question by consulting any available judicial or social scientific research on false positive rates. Or we could consider some normative principles to constrain our reasoning (for example, principles that enjoin us to be extra charitable, or extra strict, as regards the risk of falsely accusing male suspects). Or we could consider a combination of empirical and normative considerations. Ultimately, the way we arrive at this rate is not the main interest of the present project, and so for now let us assume for maximal simplicity that we do not have any information about λ and so we start with a uniform prior for λ as well (though, again, we will consider a few alternative prior specifications below). That is,

$$\pi(\lambda) \propto 1. \tag{6}$$

To repeat: we do not necessarily endorse the wisdom of using a uniform prior in the absence of information. Nor do we think that this particular example is one where we genuinely would not have any information to structure our prior. But the uniform prior has an illustrious history and many defenders in Bayesian inference (e.g. White 2010; Pettigrew 2014), and for the sake of exposition its simplicity is an asset.[8]

Now that we have a full joint prior distribution, $\pi(\lambda, \theta) \propto 1$, we can update it on the observed information (i.e. 100 thefts, 91 of which were committed by men). To do this we use a Markov Chain Monte Carlo algorithm, since it would be very difficult to calculate this by hand – in essence, we would need to consider the likelihood of the data under every possible rate of misclassification and sum over all of these possibilities. The usual approach is to use approximate Bayesian inference and estimate the posterior distribution instead. To accomplish this, we use an algorithm implemented in a general-purpose Bayesian programming language called Stan.[9] Stan estimates for us the full joint posterior distribution of λ and θ as well as the marginal posterior distribution of each. The quantity of interest to us is the mean of the marginal posterior distribution for θ, since this corresponds to the posterior credence that Jake stole the iPhone.

[8] Note: a uniform prior is equivalent to a Beta(1, 1) prior.

[9] Carpenter et al (2017). For more mathematical background see Neal (1994).

Given the evidence as stated, this probability is 0.61. (We have provided the computational details in the Appendix.) Notice that this is significantly lower than the high credence that is obtained when one blithely assumes that the prediction about Jake simply corresponds to the naked statistic – i.e. the frequency – as stated, which would be 0.91. But all that we have done is to complete the inference problem so as to account for the sort of uncertainty that would exist in realistic versions of the case. The reason the posterior credence has dropped is that we have substantial uncertainty about the error rate – λ. As λ approaches zero, with certainty, our problem reduces to Buchak's formulation of the problem and the posterior mean gradually grows closer to the frequency. For example, a Beta(1, 2) prior for λ would increase the posterior mean of θ from 0.61 to 0.74. And a very opinionated Beta(10, 30) prior for λ would increase the posterior mean to 0.87. This is to be expected since we are reducing uncertainty about the proportion of mistakes in the data.

Likewise, we can consider non-uniform priors for θ. A Beta(1, 1) prior for λ with a Beta(4, 2) prior for θ would produce a posterior mean for θ of 0.73. This is similar to the above, except that now the reduction occurs because the prior for the proportion (θ) is more opinionated – centered around 0.66 – while the error rate remains uniform, and so the data move us away from the prior, but not by a lot. We can also consider a non-uniform prior for both λ and θ. For instance, Beta(1, 3) for λ and Beta(2, 2) for θ would produce a posterior mean of 0.79.[10] And we would observe similar reductions if we restructured the parametric form of the priors. For example, a truncated normal prior with mean $\mu = 0.1$ and $\sigma^2 = 0.5$ produces a posterior mean for θ of 0.71.

Now, it might seem as though we stacked the deck here by focusing on the misclassification of men rather than that of women. One might worry that this fixes things such that it is unsurprising that the predictive probability about Jake goes down. But recall the stipulations of IPHONE THEFT: an iPhone has been stolen, and there are only two people who could possibly be the thief, one of whom is a man while the other is a woman. Now suppose that Jake is falsely accused of stealing the iPhone. This entails that Jake did not steal the iPhone. Who stole it, then? Given the setup of the case, Brenda must have stolen it. And so Jake's being falsely accused *just is* Brenda's stealing the iPhone and getting away with it — these are two equivalent descriptions of the same event. More broadly, a crime is inaccurately recorded as having been perpetrated by a man just in case it was in fact perpetrated by a woman. And this event simply *is* the event of a woman being misclassified as not having committed the crime when, in fact, she did. Thus, a false positive for one gender simply *is* a false negative for the other.[11] And so we are in fact already modeling the false negative rate for women; although we have described λ as the false positive rate for men, given the binomial nature of the case (either a man or a woman stole the phone in each instance) it can equally accurately be described as the false negative rate for women. The misclassification rates for thefts across genders cannot vary wholly independently of one another.[12]

---

[10] It is in general a good idea for the sum of the parameters in the prior beta distribution to not be too large. For example, we avoid priors like Beta(100, 200) because they would overwhelm the data and we would not learn from the evidence. This could only be sensible if we had a *lot* of prior information before encountering any observations.

[11] Here we write as if there are only two genders, following the implicit assumption in Buchak's description of the IPHONE THEFT case. But nothing changes in the model if we acknowledge that there are more than two genders; we would just say that a false positive for a man is the same event as a false negative for a *non-man*, where the category 'non-man' encompasses women and gender minorities. Alternatively, we could also model the case with multiple genders by using a multinomial likelihood with a Dirichlet prior.

[12] A false positive for a man can fail to be equivalent to a false negative for a person of another gender only if, in fact, no crime was committed at all. We think it safe to assume that cases like this, in which no theft occurs but one is somehow invented, make up only a

Moreover, our initial modeling assumptions can all be relaxed. In particular, we can further generalize our model to explicitly consider the possibility that some women are falsely convicted of stealing iPhones as well as some men. This would require an additional parameter: one that captures the false positive rate for women (which can equivalently be described as the false negative rate for men). The following figure captures this generalized situation.

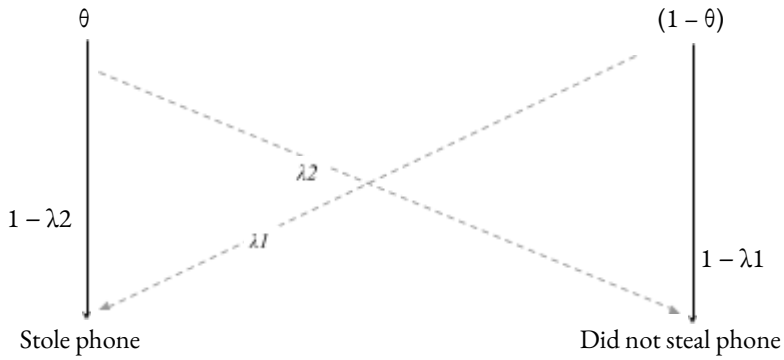

Figure 4: IPHONE THEFT with noise levels λ1 and λ1.

We now have a parameter for both types of error rate: false positives for both men and women. But now we also have to redefine φ, which we initially defined as $\theta + (1 - \theta)\lambda$ (following Equation 4). In the model with both types of error, φ is then given by the sum of $(1 - \theta)\lambda 1$ (as before) and $\theta(1 - \lambda 2)$. To avoid ambiguity, we will call this quantity φ*. That is,

$$\varphi^* = (1 - \theta)\lambda 1 + \theta(1 - \lambda 2). \quad (7)$$

The likelihood remains binomial in φ*, that is, $(\varphi^*)^x(1 - (\varphi^*))^{n-x}$. For this model, we would have to specify a prior for λ1 and a prior for λ2 as well as a prior for θ. As Babic et al (2021) explain, adding an additional error parameter – as we have just done – will further diminish the information value of the data, since there are now more ways in which individuals could have been misclassified. So, how will this affect the posterior distribution for θ? That depends quite a bit on the particular modeling assumptions – i.e., the priors assigned to both error rates as well as the original proportion. Indeed, the posterior is now even more sensitive to the prior specification, since the data have been rendered less valuable by the additional error rate parameter. In any case, if we assigned them all uniform priors as before, then the posterior credence that Jake stole the phone would diminish even more than it did previously: it would be quite close to 0.5. This is to be expected, since, with so much uncertainty about both error rates, the posterior stays very close to the prior.

---

very small proportion of false positives for any social group. (For other crimes besides theft, the proportion might be higher.) It should also be borne in mind that there can be a false negative for one group that is not a false positive for any others if a crime is committed but *nobody* is ultimately recorded as having committed it. This might quite easily happen in the case of thefts; perhaps the victim does not report the theft, or perhaps they do not press charges, or perhaps the case does not go to trial, or perhaps the jury is unable to reach a verdict and the charges are then dropped.

One might also suspect that as the evidence grows, the prediction about Jake will converge to the frequency in the long run. In other words: as we go from 100 observed thefts to 1000 observed thefts to 1 million observed thefts, one might expect that the prediction about Jake should approach 0.91. In fact, though, this is not the case. Babic et al (2021) establish that with even modest misclassification rates, there is an upper bound to the information value of noisy data. And their results are striking. For example, they demonstrate that *100,000* observations with a 30% misclassification rate are equivalent – in terms of their effect on the full posterior distribution and the associated parameter estimates – to *fewer than 4* observations with no misclassification. Moreover, this remains true even when misclassification rates are very small. For example, 100,000 observations with only a 6% misclassification rate is equivalent to roughly 100 observations with no misclassification.[13] We think it is safe to assume that we will never have 100,000 observations of stolen iPhones within a meaningful reference class. Hence, even with only very modest misclassification, *we simply may never be in a position where the evidence licenses a high credence in the predictive inference*.

One point is especially worth stressing here, since it will become important in the next section (in which we discuss implications for broader philosophical debates): our approach is entirely Bayesian. In our model, the predictive probability about Jake is not forced to be low for any non-epistemic reasons. Rather, it drops because the model now corresponds to a more realistic and complete mathematical description of the available evidence.

This really flips the traditional scholarship upside down when it comes to reasoning with statistical evidence. As we observed earlier, it is ordinarily simply taken for granted that the statistical evidence in IPHONE THEFT licenses a high credence in the proposition that Jake is the thief, and so similarly for other cases involving predictive inference. The discussion typically proceeds immediately from the statistical setup to the question of how we should resolve the apparent tension between the inappropriateness of blaming Jake (or whoever is the protagonist in the case) and the epistemic rationality of believing him to be the thief. But this is far too quick. The formal work discussed here shows that, *pace* the literature, in most realistic versions of such cases – and with any of an enormous variety of available modeling assumptions – it really is not rational to have a high credence in the proposition that the protagonist possesses the target property (e.g. that Jake is the thief). And, if one has a threshold view of belief, then unless the threshold is implausibly low it is not rational to believe this proposition either. We will expand on these points in the following, final, section.

[13] The main assumption needed for this result is that the posterior distribution can be represented in a beta form.

4. Discussion

a. Main Upshot

Our central goal in this paper has been to emphasize how much can be lost when stylized cases are used to model real-life scenarios. Toy cases are sometimes helpful, since they enable us to isolate the central parts of a problem and thus to focus the reader's attention. But all too often, and as is the case here, in abstracting away from the tricky details of real-life cases we substantively change the inference problem. It is of course possible to model IPHONE THEFT formally and to ask what credence one should have in such a case. But our point has been that the probability model for IPHONE THEFT is simply not the same as that for cases like PRISON YARD. The likelihood must be restructured to model noisy evidence appropriately. Different priors are then required, including for the added parameters, and the inference that is called for is a predictive one rather than a direct one.

This point has not been recognized in the philosophical and legal literatures on reasoning with statistical evidence. In Buchak, for instance, there is an implicit assumption that IPHONE THEFT is like PRISON YARD in all relevant respects. Buchak does consider as a potential objection to her argument the view that "statistical evidence alone cannot give rise to a high credence in certain circumstances" (p.303), which is effectively our view. But Buchak appears not to have considered the possibility that such a view can be defended on squarely Bayesian grounds; the ground that she considers for this view is just the claim "that when the proposition in question concerns the free choice of an individual and the evidence consists in the existence of an accidental correlation between belonging to a class to which that individual belongs and performing a particular act, then one should not form the credence in question" (p.304). And Buchak asserts that "[t]his proposal would be a radical revision of our theory of credence" (*ibid*.) But our view is not at all a radical revision – it's purely Bayesian. On our view, the reason that one's posterior credence in Jake's guilt should not be especially high has nothing to do with the proposition's involving the free choice of an individual and its combination with an accidental correlation about that individual's class membership. Rather, the reason is the perfectly ordinary one that the evidence does not support a high credence in this proposition. And we have simply considered the possibility of mistakes in the underlying data, which we think should be required for responsible reasoning with statistical evidence. It is unreasonable to assume *ex ante* that any data offered to us corresponds to the world exactly as it is. We must take seriously the gap between what is the case and what is observed. And that is what our model does.

In other words, our proposal simply makes use of the theoretical machinery that statisticians and formal epistemologists widely accept. The point is that when this theoretical machinery is applied to noisy, real-world cases of statistical inference, things do not look anywhere near as drastic and worrying as the stylized examples of philosophical imagination can make it appear. And we do not see this point as reflecting a deep philosophical problem with statistical evidence but rather a recognition of the full texture of the inference problem, and especially of the limited probative value of the evidence with respect to the proposition that Jake is guilty.

This means that Buchak's objection to the view that statistical evidence (sometimes) does not license high credence does not apply to our view. She writes that "[t]he primary objection to this proposal is that it severs the link between credence and rational betting behavior", reasoning that, "if only monetary gains and losses are at stake, we ought to bet that Jake stole the phone" (*ibid.*). But we certainly would not take that bet – at least not at the implied odds in light of the evidence given. And

we would indeed consider the bet under the odds implied by our own model. Hence, our view preserves the link between credence and rational betting behavior. In real-life cases of noisy data, one's betting behavior should reflect one's credences just as it usually does – and one's willingness to accept bets will reflect the fact that the evidence in such cases usually does not license a particularly high credence.

Our proposal is also distinct from some other points about the perils of statistical inference that have appeared in the recent epistemological literature. In particular, Jorgensen Bolinger (2020) and Gardiner (2020) both express doubts about whether statistical inference can be highly probabilifying, but only because both worry that such inferences can be subject to the reference class problem.[14] The worry is that all individuals belong to very many categories and that subjects performing statistical inferences typically lack justification for supposing the category at issue to be "uniquely relevant" to the probability of the individual's possessing the target property (Jorgensen Bolinger 2020, p.2429; cf. Gardiner 2020, pp.176-177). Indeed, Jorgensen Bolinger says explicitly that "[t]he justificatory hurdle can be skirted only if we lack any information about alternative classes, so that F is the relevant class simply by being the only class we have any evidence about". But our point is different from the reference class problem. As a result, our point still arises in cases in which there exists statistical evidence about just one reference class. So long as the evidence is noisy, it will not be highly probabilifying, even if no other potential reference classes are on the table. This, to us, is the particularly interesting part: it is always possible and rarely difficult to push back against a certain suggested credence by arguing that it should be based on a broader or narrower reference class, but we have explained how one ought to push back against the suggested credence in IPHONE THEFT even without questioning the reference class.

Munton (2019) explores an epistemic problem with statistical inference that is similar to, but distinct from, the reference class problem. It is that statistics can only be accurate with respect to a description of the individual entities whose properties the statistic reports, which places the individuals within a "domain", and that "the accuracy of a statistic may depend on particular idiosyncratic and contingent features of the domain", as a result of which "failing to take those contingencies into account when projecting onto a novel, unobserved instance leaves one liable to overproject" (p.231). This is a good point. But it is not our point. Munton makes clear that she is "concerned with our sense that there is something awry *even before the thinker applies the generalization to a particular individual*" (p.230, emphasis original), but she grants for the sake of argument – as do so many others in the literature – that the statistic in question is accurate with respect to the specific domain that it describes (though not with respect to that onto which it is projected). We are primarily concerned with predictive inference itself, although, like Munton, we have identified something that can go awry in agents' uncritical acceptance of statistical evidence even prior to their use of it in predictive inference. Unlike Munton's problem, however, the problem that we have described is one that arises precisely because reported statistics may *not*, in fact, be accurate even with respect to the domain to which they are circumscribed. The risk of misclassification means that we should not always assume that a reported statistic is accurate even as stated.

b. Application to Related Debates

Intuitions about statistical evidence vary in content and strength and have been put to a corresponding variety of theoretical uses. Our point in this paper is relevant to some of these intuitions and theoretical uses, but not others. Here is a quick overview.

[14] See Hájek 2007 for a thorough and influential recent discussion of the reference class problem.

i. Asymmetries between statistical evidence and eyewitness testimony

Within the legal literature, one major focus has been on a striking asymmetry: jurors are usually *less* willing to accept facts as proven based on so-called "naked" statistical evidence than based on eyewitness testimony with an equal stipulated probability of accuracy (or a lower one, even). Jurors are usually more reluctant to convict in PRISON YARD than if they are told that an eyewitness testifies that she saw Prisoner Number 37 in the yard and that this eyewitnesses has been found to be 90% reliable when it comes to identifying faces, for instance. Our point in this paper does not directly vindicate this asymmetry. But it does suggest a plausible explanation of it: the risk of misclassification is usually much more salient for statistical evidence than for stipulations about the reliability of an eyewitness. Indeed, we suspect that most jurors – and, moreover, most laypeople – either do not have a clear idea of how eyewitness reliability is assessed or have an idea that makes it seem very unlikely that it could be mistakenly assessed. (For example, people might assume that eyewitness reliability is assessed under laboratory conditions by trained professionals and therefore that it is very unlikely that their assessments involve any misclassification.) Our point in this paper also offers the *possibility* of vindication of the well-documented and widely-discussed asymmetry in trust between statistical and eyewitness evidence: if, in fact, estimates of the reliability of eyewitnesses really *do* involve little or no misclassification, then it is sensible to place more trust in them then in statistics for which the risk of misclassification is greater.

In support of our explanation, imagine that you are on a jury and are presented with the following evidence:

> This eyewitness says that she saw Prisoner Number 37 attack the guard. She has been tested and her testers recorded that 90% of the time when she says that she has seen a certain person, she has indeed seen that person, whereas 10% of the time she has in fact seen someone else. However, her testers are only human and – alas! – are subject to human error. Some proportion of the times when they record that the eyewitness correctly identifies a person will actually be times when she is incorrect. Likewise, some proportion of the times when the testers record that the eyewitness claims to have seen one person but in fact saw someone else will be times when the testers are mistaken and the eyewitness is right. Both of these error rates are unknown. What is known is just that the testers *say* that the eyewitness's reliability is 90%.

Would you think that this evidence establishes Prisoner Number 37's guilt beyond a reasonable doubt? We suspect not. Or, at least, we suspect that you would be more cagey in response to this evidence than you would have been if your evidence had consisted solely of the first two sentences of the above report. And this supports our explanation of the well-documented asymmetry: people are normally more inclined to trust eyewitness testimony than statistical evidence with an equal stipulated probability of accuracy because the risk of misclassification with respect to estimates of eyewitness reliability is rarely salient. But as the risk of misclassification *becomes* salient, the asymmetry lessens.

ii. Blanket bans on predictive inference

Some have taken the strong position that *any* increase in credence in any proposition about an individual's possessing a property based on information about the property's prevalence in a social group to which she belongs is morally unacceptable. There are various ways of arguing for this strong position, but the rough idea is typically that treating the person as a mere member of a reference class wrongs her by diminishing her individuality and her agency (as expressed in the quotation from Buchak discussed in the previous subsection, for instance). Crucially, for the strong position to follow we must also hold that a person is treated as a mere member of a reference class in this morally objectionable way whenever

anyone allows her credence in *any* proposition about the person to alter to *any* degree – even just a tiny little bit, and even if this miniscule increase does not affect her full beliefs about them – in response to information about the prevalence of the relevant property in a social group to which they belong. Our point in this paper does not vindicate this strong position, since we do not argue against predictive inference in general. And, indeed, our point in this paper undermines some arguments one might give for the strong position: one might try to argue for the strong position based on intuitions about the unacceptability of predictive inferences in cases involving noisy data, but we have provided a rival explanation of those intuitions.

We are not concerned to vindicate the strong position, however, because we doubt that it is true. The strong position is subject to a great many counterexamples; it is morally unproblematic, for instance, for a physician to update her estimate of the probability that a medical patient has a certain condition based on information about the condition's prevalence in their racial or gender group. More strongly, it seems to us that the physician's *failing* to update her credence on this available evidence might violate a duty to her patient, especially if her doing so delays their diagnosis and treatment. To us, this suggests that the strong position is too strong. Predictive inference is not diminishing or disrespectful in absolutely every case. And so more work must be done to spell out the precise conditions in which it is so. We have made no attempt to do that work here.

iii. Claims about the nature of legal proof

One further consequence of our proposal is worth noting before we end. Consider a criminal courtroom setting, wherein a defendant will be convicted only if the evidence presented at trial supports a guilty verdict beyond a reasonable doubt. Many scholars disagree on how, and indeed whether, to quantify the 'beyond a reasonable doubt' standard. Quantifying would add clarity to the law, which is a mark in its favor. And some scholars have offered sophisticated derivations of particular probabilistic thresholds based on intuitions about optimal ratios of errors produced by a trial system (though see Johnson King (2020) for an argument that such derivations fail to take seriously the possibility of misclassification). Countervailing arguments against quantifying the standard often involve cases of statistical evidence: for instance, in PRISON YARD, if the standard of proof is set at 0.9 then any randomly-selected prisoner – or indeed all of them – could be found guilty beyond a reasonable doubt, despite the fact that ten must be innocent. And with higher probabilistic thresholds, alternate versions of the case with different proportions of actually-guilty prisoners yield the same result. Scholars who are skeptical of probabilifying evidentiary burdens worry about cases like this, in which it seems all-too-easy to satisfy a high probabilistic threshold on the basis of evidence that does not strike them as sufficiently narrowly tailored to the individual defendant.

In many, although not all, such cases, our view provides an alternative explanation of the unease one might feel at the prospect of a conviction based on statistical evidence. Let's start with more realistic cases – those like IPHONE THEFT. Suppose that we set the burden for convicting Jake of theft at 0.85; very few people would think that 'beyond a reasonable doubt' can correspond to a lower probability than this, so this is a fairly conservative threshold view of the relevant evidentiary burden. But the statistical evidence presented in court would ordinarily involve some risk of misclassification. And so, given our discussion in the previous section and given Babic et al (2021)'s asymptotic results, it is entirely possible that in many proceedings *no amount of statistical evidence would ever get us above the requisite threshold* of 0.85! This offers a novel diagnosis of what is wrong with statistical evidence in courtrooms: instead of assuming that such evidence leads to a high predictive probability and trying to identify a problem elsewhere, in most realistic cases we can simply point out that the noisy data do not in fact yield a high posterior probability of guilt. Hence our main point in this paper,

concerning the impact of the risk of misclassification on predictive inference, is highly significant for those using intuitions about statistical evidence to defend conclusions about the standard of proof. The difference between credences of 0.91 and 0.61 in the proposition that Jake stole the iPhone is the difference between a credence that many legal scholars have proposed as the probabilistic threshold equivalent to proof beyond reasonable doubt[15] and one that is plainly far too low to meet this standard.

This diagnosis of what is wrong with statistical evidence in courtrooms is much simpler than some others in the literature. For instance, as we noted in section 2, Buchak introduces belief-credence dualism and the claim that legal verdicts can only be based on full beliefs in order to explain the intuition that there is something amiss with believing Jake guilty based on the evidence in IPHONE THEFT, and Moss introduces the idea of probabilistic knowledge and the claim that legal verdicts must be based on knowledge of probabilistic contents in order to explain the same intuition. Our alternative diagnosis is a great deal simpler than these two, since it requires no additional theoretical machinery beyond the usual apparatus of Bayesian predictive inference.

It must be acknowledged that there are some limitations to this alternative diagnosis. Importantly, if we really are in a case like PRISON YARD – where the evidence is unambiguous, the prisoners are fully interchangeable, and the misclassification rate is *ex ante* known, with certainty, to be zero – then our model would agree with the ordinary setup and imply that the probability that any randomly-selected prisoner participated in the attack corresponds to the proportion of prisoners in the yard who were participants. Scholars who are skeptical of probabilifying evidentiary burdens can continue to make their case based on examples like these, in which their skepticism remains intuitive. And fictional examples suffice for them to make their points. Still, cases like this are vanishingly small in real life, since perfect interchangeability and known-to-be-zero misclassification rates are generally works of theoretical fiction. So, whatever our best scholarship suggests is wrong with statistical inference in toy cases like PRISON YARD, in real-life cases there is often a much, much simpler problem with such an inference: the evidence does not support the conclusion.

[15] The *locus classicus* here is Kaplan (1968). See also the survey of federal judges in McCauliff (1982); McCauliff found that judges tend to associate 'beyond a reasonable doubt' with a probability of 0.9.

Appendix.

To obtain the posterior marginal estimate for θ in Section 3, we use the following code. The model is written as a .stan file, and it is then implemented in R.

```
\\ saved as ose.model1.stan
data {
int < lower = 1> n;
int < lower = 1, upper = n> Y;
}

parameters {
real < lower = 0, upper = 1 > lambda;
real < lower = 0, upper = 1 > theta;
}

transformed parameters {
real < lower = 0, upper = 1 > tau;
tau = theta+(1-theta)*lambda;
}

model {
Y ~ binomial(n, tau);
lambda ~ beta(1, 1);
theta ~ beta(1, 1);
}
\\
model_path <- "ose_model1.stan"
model_mme = stan_model(model_path)
stan_data <- list(Y = 91, n = 100)
fit_main <- sampling(model_mme, data = stan_data,
warmup = 10000, iter = 100000, chains = 2, cores = 1,
thin = 1, control = list(adapt_delta = 0.99, stepsize = 0.001, metric = "dense_e"))
print(fit_main)
```

To obtain the posterior marginal estimates for θ when there are two misclassification parameters, we use the following code instead.

```
\\ saved as ose.model2.stan
data {
int < lower = 1> n;
int < lower = 1, upper = n> Y;
}

parameters {
real < lower = 0, upper = 1 > lambda1;
real < lower = 0, upper = 1 > lambda2;
real < lower = 0, upper = 1 > theta;
}

transformed parameters {
real < lower = 0, upper = 1 > tau;
tau = theta*(1- lambda2)+(1-theta)*lambda1;
}

model {
Y ~ binomial(n, tau);
lambda1 ~ beta(2, 18);
lambda2 ~ beta(2, 18);
theta ~ beta(1, 1);
}
\\
```

The basic code for executing the model remains unchanged.

This completes the Appendix.

References.

Buchak, Lara (2014). "Belief, Credence, and Norms." *Philosophical Studies* 169:285–311.

Babic, B., Gaba, A., Tsetlin, I., & Winkler, R. (2021). "Normativity, Epistemic Rationality, and Noisy Statistical Evidence." *The British Journal for the Philosophy of Science*.

Carpenter, B., Gelman, A., Hoffman, L. M., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Stan, A. R. (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, 76(1), 1–32.

Cheng, E. K. (2013). "Reconceptualizing the Burden of Proof." *The Yale Law Journal*, 122(5), 1254-1279.

Cohen, L. Jonathan (1977). *The Probable and the Provable.* Oxford: Clarendon Press.

Colyvan, Mark, Helen M. Regan and Scott Ferson (2002). "Is it a Crime to Belong to a Reference Class?" *The Journal of Political Philosophy* 9(2):168-181.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). "Hybrid Monte Carlo." *Physics Letters B*, *195*(2), 216–222.

Enoch, David, Levi Spectre and Talia Fisher (2012). "Statistical Evidence, Sensitivity, and the Legal Value of Knowledge". *Philosophy and Public Affairs* 40(3):197-224.

Gaba, A. (1993)." Inferences with an unknown noise level in a Bernoulli process." *Management Science*, *39*(10), 1227–1237.

Gaba, A., & Winkler, R. L. (1992). "Implications of errors in Survey Data: A bayesian model." *Management Science*, *38*(7), 913–925.

Gardiner, Georgi (2020). "Profiling and Proof: Are Statistics Safe?" *Philosophy* 95:161-183.

— (forthcoming). "Legal Evidence and Knowledge." Forthcoming in Maria Lasonen-Aarnio and Clayton Littlejohn (eds.), *The Routledge Handbook of the Philosophy of Evidence*.

Hájek, Alan (2007.) "The Reference Class Problem is Your Problem Too." *Synthese* 156:563-585.

Hawthorne, John. (2004). *Knowledge and Lotteries.* Oxford: Oxford University Press.

Johnson King, Zoë (2020). "The Trouble With Standards of Proof". *Synthese* 199(1-2):141-159.

Johnson King, Zoë, & Babic, Boris (2020). "Moral obligation and epistemic risk." In M. Timmons (Ed.), *Oxford Studies in Normative Ethics Volume 10*, 81–105.

Jorgensen Bolinger, Renee (2020). "The Rational Impermissibility of Accepting (Some) Racial Generalizations". *Synthese* 197(6):2415-2431.

Kaplan, J. (1968). "Decision Theory and the Factfinding Process." 20 *Stanford Law Review* 1065.

Kyburg, H. E. (1961). *Probability and the logic of rational belief.* Middletown, CT: Wesleyan University Press.

Kyburg, H. E. (1974). *The Logical Foundations of Statistical Inference.* Dordrecht: Reidel.

Littlejohn, Clayton (2020). "Truth, Knowledge, and the Standard of Proof in Criminal Law". *Synthese* 192(12):5253-5286.

Levi, Isaac (1977). "Direct Inference". *The Journal of Philosophy* 74 (1), 5–29.

McCauliff, C. M. A. (1982). "Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitutional Guarantees?" 35 *Vanderbilt Law Review* 1293.

Moss, Sarah (2018). *Probabilistic Knowledge.* Oxford: Oxford University Press.

Munton, Jessie (2019). "Beyond Accuracy: Epistemic Flaws with Accurate Statistical Generalizations". Philosophical Issues 29(1):228-240.

Neal, R. M. (1994). "An improved acceptance procedure for the hybrid Monte Carlo algorithm." *Journal of Computational Physics*, *111*(1), 194–203.

Nesson, Charles (1979). "Reasonable Doubt and Permissive Inferences: The Value of Complexity". *Harvard Law Review* 92:1187–1225.

Nesson, Charles (1985). "The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts." *Harvard Law Review* 98(7):1357-1392.

Pardo, Michael (2018). "Safety vs. Sensitivity: Possible Worlds and the Law of Evidence." *Legal Theory* 24(1):50-75.

Pettigrew, Richard (2014). "Accuracy, Risk and the Principle of Indifference." *Philosophy and Phenomenological Research* 92(1):35-59.

Pritchard, Duncan (2018). "Legal Risk, Legal Evidence and the Arithmetic of Criminal Justice." *Jurisprudence* 9:108–119.

Pundik, Amit (2008). "Statistical Evidence and Individual Litigants: A Reconsideration of Wasserman's Argument from Autonomy." *International Journal of Evidence and Proof* 12:303-331.

Tribe, Lawrence (1971). "Trial By Mathematics: Precision and Ritual in the Legal Process." *Harvard Law Review* 84(6):1329-1393.

Thomson, J. J. (1986). "Liability and Individualized Evidence." *Law and Contemporary Problems* 49(3): 199-219.

Wasserman, David T. (1991). "The Morality of Statistical Proof and the Risk of Mistaken Liability." *Cardozo Law Review* 13 935:942-43.

Winkler, R. L., & Gaba, A. (1990). "Inference with Imperfect Sampling from a Bernoulli Process." In N. Longford et al. (Eds.), *Bayesian and Likelihood Methods in Statistics and Econometrics* (pp. 303–317). North-Holland.

White, R. (2010). "Evidential Symmetry and Mushy Credence." In Tamar Szabo Gendler and John Hawthorne (Eds.), *Oxford Studies in Epistemology Volume 3*, 161–186.