Explaining the Justificatory Asymmetry between Statistical & Individualized Evidence[*]

Renèe Jorgensen Bolinger[†]

ABSTRACT: In some cases, there appears to be an asymmetry in the evidential value of statistical and more individualized evidence. For example, while I may accept that Alex is guilty based on eyewitness testimony that is 80% likely to be accurate, it does not seem permissible to do so based on the fact that 80% of a group that Alex is a member of are guilty. In this paper I suggest that rather than reflecting a deep defect in statistical evidence, this asymmetry might arise from a general constraint on rational inquiry. Plausibly the degree of evidential support needed to justify taking a proposition to be true depends on the stakes of error. While relying on statistical evidence plausibly raises the stakes by introducing new kinds of risk to members of the reference class, paradigmatically 'individualized' evidence—evidence tracing back to A's voluntary behavior—can lower the stakes. The net result explains the apparent evidential asymmetry without positing a deep difference in the brute justificatory power of different types of evidence.

1 The Asymmetry

Here's a simple picture: something is evidence for p just if it raises the epistemic probability that p, and the strength of evidence— its power to justify epistemic actions like inference, belief, etc.—can be read directly off of these probabilities: e_1 is stronger evidence for p than e_2 iff the probability of p given e_1 is greater than the probability of p given e_2. Let's call this SIMPLE-PROBABILISM. This picture does not distinguish between statistical information and direct observation when thinking about their power to ground inferences. But sometimes there is an apparent asymmetry in their justificatory power; for example, consider two cases:

(a) PRISON YARD-A[1] – One hundred prisoners exercise in the prison yard. Ninety-five prisoners together assault a guard. Security footage reveals five prisoners standing against the wall refusing to participate, but is too grainy to identify them. The guard cannot identify his assailants. Alex is prosecuted for the assault, on the grounds that we know he was one of the prisoners in the yard, and therefore we can be 95% confident that he participated in the assault.

(b) PRISON YARD-B – One prisoner attacks a guard, and there's an eyewitness whom we know to be accurate under similar conditions 85% of the time. Alex is prosecuted on the grounds that the eyewitness testifies that Alex was the assailant.

Let p be the proposition that 'Alex participated in the assault'. The following three judgments seem true:

1. The evidence in the (a) case renders p more probable than the evidence in the (b) case.
2. In the (a) case, our evidence does not suffice to establish p.
3. In the (b) case, our evidence does suffice to establish p.

This puzzling trio of judgments is an instance of the *statistical evidence proof paradox*.[2] Given SIMPLE-PROBABILISM,

[*]To appear in *Truth and Trials: Dilemmas at the Intersection of Epistemology and Philosophy of Law*, Edited by Zachary Hoskins and Jon Robson (University of Nottingham)

[†]My thanks to Christian Barry, Geoff Brennan, John Broome, Maegan Fairchild, Philip Pettit, Wlodek Rabinowitz, Jeremy Strasser, Katie Steele, and James Willoughby, as well as to the audience at the 2017 Philosophy & Economics Workshop at Australian National University, and the participants at the 2019 Philosophy of Statistics Summer Seminar at Virginia Tech for fruitful conversations on earlier versions of this material.

[1]The PRISON YARD case was introduced by Nesson (1979, 1193) and has been widely discussed among legal evidence theorists.

[2]See Cohen (1977), Redmayne (2008), and Pardo (2019) for illuminating and more comprehensive discussion of the various proof paradoxes.

(1) implies that the evidence for p in the (a) case is stronger than in the (b) case, but (2) and (3) jointly seem to imply the opposite.

Some take this puzzle to motivate a position I'll call STATISTICS-DEFICIENCY: when e is statistical evidence, it may be able to contribute to an agent's rational credence in p, but it cannot justify all-out attitudes like full belief, constitute knowledge, or license the agent to assert or issue a judgment that p.[3] That is, that (1)-(3) are all true, because there is a justificatory asymmetry between statistical evidence and other kinds of evidence due to a particular defect in statistical evidence. The canonical source for this skeptical stance toward statistical evidence is Laurence Tribe (1971), who included all explicitly probabilistic evidence in the scope of his criticisms.[4]

Advocates of the STATISTICS-DEFICIENCY have offered a wide range of explanations for it.[5] Colyvan, Regan, and Ferson (2001) stress the difficulty of establishing an appropriate reference class, as well as the tenuous connection between statistical evidence and the agency of the subject of the trial. They consequently praise Judge Newman for vacating a sentence (in *United States v. Shonubi*)[6] that had been based partially on statistical evidence, on the grounds that "it did not constitute specific evidence".

Though it is widely held, not everyone shares this view. Peter Tillers (2005) is sharply critical of both Judge Newman and Colyvan, Regan, & Ferson's discussion of the case. He levels two main objections. First, that "the attempted distinction between specific and non-specific evidence is almost unintelligible" (Tillers, 2005, 36), and so the requirement that evidence be specific would preclude all group-to-individual inference, including inferences from the person's own past behavior. Second, that the exclusion of statistical evidence was improperly justified, because it "is best explained or justified by considerations of justice or social policy rather than by epistemic, inferential, or statistical principles". He contends that it is inescapable that individuals will be saddled with "a vast multitude of evidentiary signs that are generated by the cosmos and other people, rather than by the individuals about whom the inferences are to be made" and thus the specific evidence requirement is wishful

[3] Buchak (2014); Jackson (2018) deny that evidence couched in statistical terms can justify full belief; Moss (2018) deny that it grounds knowledge; Thomson (1986) holds that it does not license assertion.

[4] Tribe is clear that his objection to statistical evidence is not simply that it is probabilistic in nature, rather than delivering certainties: "I am, of course, aware that all factual evidence is ultimately "statistical," and all legal proof ultimately "probabilistic," in the epistemological sense that no conclusion can ever be drawn from empirical data without some step of inductive inference". . .My concern, however, is only with types of evidence and modes of proof that bring this probabilistic element of inference to explicit attention in a quantified way." (Tribe, 1971, 1330, fn 2)

[5] As a representative sampling: Enoch, Spectre, and Fisher (2012) argue that evidence must be *sensitive*: for e to justify belief in p, it must be that if p weren't the case, we wouldn't have e. This is a modal notion, and evidence is stronger the larger the range of possible worlds in which this counterfactual holds. Statistical evidence often fails this constraint, since the generalization will hold even if the particular individual on trial doesn't have the property in question (though not always; see criticisms in Blome-Tillmann (2015), Littlejohn (2017) and Gardiner (2018)). Pritchard (2017) argues that evidence must be *safe*, rather than risky: for e to justify belief in p, it must not be easily consistent with $\neg p$. Roughly, given that we have e, it must be true that p in all the nearby possible worlds. Smith (2017a, 2017b) holds that evidence must provide *normic support* for the proposition it justifies, which e does for p just if, given that e is true, if p fails to be true we'll need some explanation to reconcile the two facts; it's not normal, in an important sense, for e to be consistent with $\neg$P. Risinger (2004, 2018)'s 'surprise' theory, on which e is good evidence for p if, given e, we would be surprised to learn that p is false, can be read as a variant of the normic support model as well. Haack (2012) (as well as Brennan-Marquez (2017) and others) suggest that evidential strength amounts to explanatory power and coherence, whereas statistical generalizations typically speak to probabilities without explaining why they should hold, and so are relatively weak evidence.

[6] 103 F.3d 1085 (2d Circuit, 1997)

thinking, driven by "a moral yearning for perfectly equitable and just evidence and inference"—but wishing can't make it so.

I'll largely set aside the challenge of articulating what, precisely, the requirement that evidence be 'specific' or 'individualized' amounts to, in order to focus on the second half of Tillers' complaint. I will argue that there is a straightforwardly epistemic reason to think that statistical evidence will generally have little probative value, and so be less valuable than other sorts of evidence. Additionally, there is an appropriate role for political or distributive considerations to justify the exclusion of certain forms of statistical evidence, even if it is impossible to limit our inferences to just the individual's own behavior. Even to pursue this more limited goal, we'll need some idea of what the difference is between statistical evidence and other forms.

Both individualized and statistical evidence can be merely probabilistic. The distinctive feature of statistical evidence is that it leverages population-level generalities in very particular ways—but there are two variations we should distinguish from each other. The first, *Statistical Inference*, is inductive: it uses data gleaned from an observed sample to build a statistical model in order to estimate frequencies in a target population, but makes no claims about specific individuals within them.[7] The second, *Actuarial Inference*, is a broadly deductive method for using information about population-level frequencies to set your degree of confidence in claims about an individual member of that population. As a first step, it assumes that your credence in an arbitrary member of G's having a property F should match the population-level frequency of having F among G-members. Then it advises that, in the absence of better information about a particular member a of G, your confidence that a *has* F (which I'll abbreviate Fa) should match the probability of Fx for an arbitrary x in G. This form of reasoning is also called *direct inference*, and it is this that underwrites the inference in PRISON YARD A.

These two methods raise slightly different questions. It is relatively easy to distinguish model-based statistical inference from other forms of reasoning, and to explain why the epistemic justification it yields might be fragile and constrained. There are a host of questions we should ask about the adequacy of the model, the representativeness of the sample population, the accuracy of the measure for the target property, the power of the test to detect the measured property, etc. But the main arguments against using 'statistical evidence' — including Tribe (1971)'s own original complaint — do not raise concerns about any of these issues. They are content to assume that the statistical inferences are correct, and the resulting estimates are reliable, even reflecting the true frequency of the trait in the target population. Forbidding reliance on statistical inference will not secure the prescription critics of statistical evidence take their arguments to establish, namely that (as Duff, 1998, puts it) when judging a person's guilt, "we must rely only on evidence related to him as an individual agent, not on evidence related to him only as a member of an actuarial group." The central question concerns only the actuarial inferences: if we discover that G has a higher than average rate of Fness, *what follows from knowing that a belongs to G*? Is it evidence that Fa?

There is an *intuitive* difference between actuarial inference and 'specific' evidence, roughly tracking the difference between particular observations and broad generalizations, but it doesn't hold up well under scrutiny. As Schauer (2003, 103) points out, "even the processes that initially appear to us to be 'direct,' 'actual,' or individualized turn out to rely far more on generalizations from past experience than is often appreciated." The

[7] Importantly, this mode of inference makes predictions about the distributive properties of a trait or event in a population, within certain confidence intervals; it does not make claims about the properties of specific individual members.

inferences that allow us to leverage observations about someone's appearance or behavior to draw conclusions about criminality rely on generalizations (e.g. 'most of the time someone weaving between lanes in such a way is intoxicated') as a suppressed premise. Consider eye-witness testimony, a form of paradigmatically individualized evidence. A witness asserts that she saw someone with salient features in common with the defendant (height, weight, hair color, perhaps driving a car that looks like one owned by the defendant, etc.) at the scene around the time of the crime. Since it's unlikely that she saw someone who had all those properties and was not the defendant, we conclude that the witness saw the defendant. But this final bit of reasoning is what makes the eyewitness's testimony evidence *about* the defendant. When this premise is made explicit, as it was in *People v. Collins*,[8] it is readily recognized and dismissed as statistical. So, the difference cannot be whether the evidence depends on a generalization at any point.

In the absence of a consistent, principled articulation of the division, proposals to exclude statistical evidence are vulnerable to the charge that they are at best just an *ad hoc* collection of specific prohibitions, which are not epistemically motivated. So even once we've narrowed the scope of the complaint from being against *all* probabilistic inferences to only an objection to actuarial inference, we still owe an answer to the challenge: why exclude *any* evidence that might bear on our inquiry? Won't we be more accurate in each particular case if we allow ourselves to draw on all the relevant information, statistical or not? Sometimes the generalizations will associate various forms of criminality with protected classes, and this may seem unfair or unjust, but if we bar reliance on those grounds, isn't this—as Tillers says—simply allowing our moral discomfort with the truths to lead us to engage in wishful thinking?

2 Explaining the Asymmetry

This is a question which can be posed at the individual level as well as the institutional. I have discussed why it is rationally impermissible for individuals to rely on particular generalizations in earlier work (Bolinger, 2018); this paper extends the analysis to institutional contexts. I contend that the apparent justificatory asymmetry is both real and grounded in properly epistemic considerations.

The general form of my argument is this: the nature and aims of rational deliberation generate a constraint on inquiry that is especially difficult for actuarial information to satisfy. As a consequence, the amount of evidential support needed clear the justificatory threshold for closing inquiry, when relying on actuarial inference alone, can be a good deal higher than when basing a judgment on other forms of evidence. This explains the observed justificatory asymmetry. In closing, I'll suggest that given its low epistemic value, and the social and moral costs of having a policy that allows the use of actuarial evidence, we have most reason to exclude reliance on these forms of evidence as a matter of public policy.

2.1 rational settling

All sorts of inquirers—not just individuals, but fact-finding bodies, and legal courts—can be epistemic agents. In a trial, the finder of fact aims to determine whether there is sufficient evidence to establish a specific proposition, namely, that the defendant is guilty (or liable), so that they can decide whether to sentence or acquit. We can get insight into the rational constraints on this institutional inquiry by paying careful attention to the structure of rational inquiry in the abstract, borrowing a bit from formal epistemology.

[8] 68 Cal.2d 319, 66 Cal.Rptr. 497 (1968)

An ideal epistemic agent represents all the possible ways the world might be, given their evidence, in a theoretical construct called a *state space*. The intersections of propositions define regions of the space, and the rational agent assigns each region a degree of confidence according to its probability of being true given their evidence. An ideal and unlimited agent should only represent certainties as true throughout the space (receiving credence 1), and only their negations as false (credence 0). Anything that *could possibly* be false (but is not certainly so), on her evidence, should receive a credence greater than 0 and less than 1. When such an agent deliberates about what to do, they should weight the possible outcomes of the actions they consider by their relative credence in each possibility, and then follow an appropriate decision rule.[9] Since a full state space forms a complete partition of logical space—every *evidentially possible* scenario receives some positive non-zero credence—the number of possibilities represented and considered in this way will be vast.[10] But that is no problem for an idealized agent with limitless epistemic resources.

Of course real agents, whether individual or institutional, have limited epistemic resources, and some possibilities are of greater consequence than others. Since a cognitively bounded agent can track only a small finite number of possibilities, they must focus on the ones that make a difference to how it is rational for them to act. So, even if her maximally careful credences in any possibility remain as represented in the full statespace, a rational *but limited* agent will have to work with a simplified space, which I'll call the *active frame*, representing only the propositions of interest, when deliberating about what to do. Several propositions that are not logical truths—and in which the agent's degree of confidence, given their evidence, is noticeably less than 1—will be simply taken for granted in the active frame. If p is one of these propositions, all of the regions of the frame will assume that p holds. The frame needn't explicitly represent p as true; it just won't have a region representing the possibility that $\neg p$. I will say that such a proposition is *settled*. To settle that p is to consider the question of *whether* p closed: to stop actively seeking evidence bearing on it, stop worrying about the risks associated with the possibility that $\neg p$ when deciding how to act, etc. It does not necessarily involve becoming certain in p.

The difference between keeping track of a proposition and settling it shows up when the bounded rational agent deliberates about what they should do. If they're keeping track of *whether* p in their active frame, then when they face a p-dependent decision, they'll draw up a decision table comparing the outcomes of each alternative under the assumption that p is true with their outcomes under the assumption that $\neg p$. Then, using that information as an input, they'll select the option that fares best according to their preferred decision rule for risky tradeoffs (e.g., maximize expected utility). By contrast, if they've already settled that p, they'll just operate under the assumption that p when selecting their action; they'll never draw up a column for the possibility that $\neg p$.[11] Legal findings of fact and judgments of guilt are not hedged bets; rather, they are institutional correlates

[9] The question of *which* decision rules might be appropriate to legal judgments will take us too far afield, so I set it aside.

[10] Such a partition may well represent as possible propositions that are, in fact, logical impossibilities; an agent whose evidence does not establish a particular identity fact (e.g. a=b) will have various logical impossibilities (including a≠b) in her logical state space. We will want to say something similar for logical theorems that are difficult to recognize, or conjectures that if true, are necessarily true, but accommodating these presents a more difficult technical challenge, which I will set aside for this paper.

[11] I of course do not mean to suggest that real agents pause to literally draw up decision tables when deliberating. Some people do, but I take it that most instead employ heuristics that roughly approximate this reasoning structure. What I've called 'settling' maps pretty well onto an epistemic state variously called 'acceptance' (Bolinger, 2018; Bratman, 1992), 'belief' (Ross & Schroeder, 2014), and 'premising' (Locke, 2013). Whether this attitude is appropriately called *belief* in the individual case is a question that divides theorists; I'm not particularly fussed about what we call it. Our interest is only in the fact that at the institutional level, something structurally like *settling that p* is a

of all-out individual states like belief, assertion, or premising. I do not suggest that they *are*, or even necessarily directly involve, belief; but simply that they are ways of *settling a proposition*.

Which propositions are worth tracking changes over time for an agent, as their information, values, and available actions change.[12] In settling that p, the agent moves from a frame that has regions encoding information about the evidential probability of $\neg p$ to one that doesn't; and from a degreed attitude concerning p to a binary one. Making this transition involves information loss. The main reason to do it is that it simplifies their reasoning, freeing up epistemic resources for other tasks, including tracking other propositions. But because it discards their recorded degree of credence in $\neg p$, it can only be rational when that information doesn't earn its keep: when the value of tracking p is not sufficient to justify the epistemic cost of doing so.

The value of tracking p is a function of (i) what difference the remaining uncertainty about p makes to what action is most rational for the agent to perform, and (ii) how much they can expect their assessment of p to change on new evidence. If an agent's current evidence renders their remaining uncertainty about p irrelevant to what they should rationally do, they are *practically adequate*.[13] If their evidence grounds a stable credence that is unlikely to change much as they gather more information, their credence is *resilient*. Putting all of this together yields the following as a constraint on rational settling:

> **Rational Settling Constraint**: It is rationally permissible for an agent S to settle that p on evidence e only when their credence in p on e is *resilient* and *practically adequate* for all their anticipatable p-dependent decisions.

This constraint entails that whether an agent can rationally settle that p on her current evidence cannot be simply read off the evidential probability of p. The justification for the rational settling constraint arises from considerations about the sort of epistemic process *settling* is; it is peripheral whether the agent doing the settling is an individual or an institution, and it does not matter what the subject matter is. So if issuing a legal judgment is a form of settling, then we should expect it to be governed by this constraint. Despite the initial air of paradox around cases like PRISON YARD, if something like the *Rational Settling Constraint* holds, we shouldn't be surprised that we can find case pairs in which whether evidence is sufficient to settle p does not directly track how probable it makes p. I maintain that the observed asymmetry between the justificatory force of individualized and actuarial evidence is due to the fact that the latter has a harder time satisfying the constraint. More specifically, the asymmetry arises from two facts:

1. the population-level generalizations underwriting actuarial inference are a less resilient basis for credence than paradigmatically individualized information, and
2. it is difficult for actuarial evidence to render an agent practically adequate.

Let's take these one at a time.

prerequisite to issuing a legal judgment that p.

[12] To formally model such agents, we'll have to allow for both the introduction of some propositions that aren't yet in their active frame, and elimination of some that propositions that are. I won't pursue the former project here; I am interested only in the constraints on when agents can discard possibilities from the set they're keeping track of.

[13] Anderson and Hawthorne (2019) give the following, more technical definition of practical adequacy: "S is practically adequate with respect to p iff the top-ranked element(s) in S's actual preference ranking do not differ from the top-ranked element(s) in her ranking conditional on p."

2.2 resilience

Rather than tracking the conditional probability of p on e, or the total amount of evidence amassed, resilience is a measure of how stable a credence is: how little variance we should expect as the agent acquires additional evidence. In institutional contexts, it corresponds roughly to some of the ways that legal theorists unpack the notion of 'evidential weight'.[14] Credences based solely on reference-class based frequencies or population-level propensities tend not to be very resilient, especially as compared with those based on 'specific' or 'individualizing' evidence. This is because the necessary inferential premise—that p is as likely to be true of the particular individual before us, Alex, as of any arbitrary member of that class—can easily be disrupted by new evidence.

There are many ways that could happen. One is by learning that Alex is a member of another, plausibly relevant reference class, with a very different conditional probability for p. We could learn that Alex is a pacifist, and that only 5% of pacifists ever participate in assaults. This would be new evidence about a competing reference class, and on learning it, we can no longer simply rely on the probabilities from the 'prisoners-in-the-yard' class.[15] It's also possible for population-level correlations to reverse as you look at increasingly specific subsets of the class. This is nicely illustrated in the 1973 fall admissions for the University of California, Berkeley. At the university level, the probability of being admitted was much higher for men (44%) than women (35%). But at a department-level this trend reversed: women were being admitted at a higher rate than men; they were simply applying in larger numbers to the highly selective departments, while the opposite was true of men.[16] Both of these scenarios involve acquiring actuarial evidence that conflicts with our initial information. But even without such evidence, credences about any particular individual's traits are quite fragile when based on a group-level frequency fact. Any additional evidence we have about Alex that differentiates him from an arbitrary member of the group will carry more information about him than the population-level generalization did, and so should have accordingly more influence on our credences.

For these reasons, it is difficult—though not necessarily impossible—for actuarial evidence to satisfy the resilience requirement of the RATIONAL SETTLING CONSTRAINT.

2.3 practical adequacy

The practical adequacy constraint is slightly more involved. Whether some evidence e renders an agent practically adequate has a decision-theoretic structure: it's a function of the evidential probability of p and the stakes of error for foreseeable p-dependent decisions. The more costly it would be to mistakenly act on the assumption that p, the more sure we'll need to be that p in order to achieve practical adequacy. For legal judgments, the most salient cost of error is wronging the immediate subject of the trial by treating them as guilty or liable when they are not.

---

[14] Much of the discussion of evidential weight in Cohen (1977, 271) appeals to the importance of the expected stability of one's credences–which is resilience. However, as Cohen developed the notion, *weight* refers to a proportion or amount of total evidence, while resilience is better understood as how difficult it would be for new evidence to substantially shift one's credences. Several contemporary proposals for a comparative plausibility standard of evidential strength (made as part of a general move toward explanationism and away from simple probabilism are articulated in ways that strongly evoke resilience (see Allen & Pardo, 2019, for a helpful overview and summary of many of these proposals) .

[15] This was in fact one of the main weaknesses in statistical evidence that Colyvan et al. (2001) emphasize.

[16] This phenomenon, Simpson's Paradox, is one of a family of structurally similar cases known as the 'ecological fallacies'.

Settling that p on actuarial grounds runs a very particular sort of risk: not simply that we will impose the costs of a false finding on Alex, but that we will do so *because of* his membership in the group that forms the basis of the generalization. I will assume (solely for simplicity) that there is no moral disvalue to imposing this cost when p is actually true; that the prisoners who participated in the assault are not wronged by facing a higher risk of being found guilty were they (counterfactually) innocent. Still, actuarial evidence consists in a generalization about the frequency of a feature of interest in a group defined by a *different* property, G. Non-homogenous groups have *some* group-members of whom p is false, whom we would wrong were we to treat them as if p. Though only one individual is directly at risk in any given trial, if the property determining membership in the G-class is relatively stable, then having a policy of taking *being* G as evidence of p entails that simply being G increases a person's odds of being falsely convicted *if* they are tried.

How serious this cost is will depend on what sort of property G is: how difficult it is to avoid, how stable or visible it is, whether it is chosen, and perhaps how central to autonomy. Individuals have a particularly strong moral and justice-based complaint against being subjected to heightened risk of false conviction *just because* of an unchosen, identity-tracking property (race, height, gender, etc.). They may have a similarly strong complaint against facing an increased risk on the basis of a chosen social identity to which they are morally entitled, or which is central to valuable exercise of their autonomy (e.g. religion). So if G is the sort of property that individuals cannot avoid having, or that they *ought* to be free to have without facing extra risks of p-based error, there is a high cost of error associated with using statistics drawn from the G-class as evidence for p. When G is avoidable, or is not a moral entitlement (gang membership, wearing a uniform, etc.), the cost of error is less severe. The original PRISON YARD case takes *presence in the yard* as the class-defining property; this is not identity-tracking, but if the prisoners' movements are controlled such that they cannot leave the yard at will, it is not avoidable, either.

The more stable or enduring the reference class, the more relying on it will concentrate the risk-exposure. So, the more difficult or costly it is for individuals to leave the reference group (by no longer having G), and the more often that group is used in judgments of a p-type, the greater the burden of the risk imposed, and hence the greater the moral wrong done in imposing it. Conversely, generalizations that draw on fragile or ad-hoc groups (e.g. the set of individuals matching a very specific eye-witness description) tend not to repeatedly subject the same set of members to risk of error, and so are easier to justify using, than generalizations that draw on more stable social categories. Less specific eye-witness descriptions are more complicated; I will return to them at the end of §3.

All else equal, then, it will often be a bit more difficult to achieve practical adequacy when relying on actuarial evidence. But often 'all else' is not equal: the forms of evidence that are paradigmatically individualized—confessions, observed behavior, and eye-witness testimony—also can have a stakes-lowering effect, and so make it easier to achieve practical adequacy. To see this, consider a third variant of the PRISON YARD case. Holding fixed that Alex is actually innocent, imagine that he freely submits a signed confession to participating in the assault. It no longer seems that he can legitimately complain against our concluding that he is guilty, given that he told us that he is. This holds even though Alex suffers some costs from being mistakenly treated as guilty.[17] The best explanation is that something like *Responsible Claim Mitigation* is true: in asserting that p, Alex undermines

[17] It is important here that the confession be *freely* given; it is likely that confessions obtained in coercive or intimidating plea-bargain contexts will not have the same effect.

his complaint against the risk of our making a p-based error.[18]

> *Responsible claim mitigation*– A's responsible performance of an *avoidable* behavior b with the meaning that p mitigates A's complaint against S's settling that p, and hence weakens A's complaint against the costs of a p-based error.

I contend that agents' moral complaints against suffering the costs of a p-based mistake are mitigated when they avoidably and responsibly supply the evidence that grounds the error. If something of this sort is true, we might get a scale of avoidable behaviors that are variously strong evidence for guilt, that the agent can be morally expected to avoid performing, when innocent. Direct observation of such behavior has a stronger effect than testimony from an eyewitness to such behavior, but perhaps both function to lessen the agent's complaint against the risk of false conviction. To the extent that our judgment whether p rests of evidence of this kind, it is easier to achieve practical adequacy, because the wrongs risked are less weighty.

Depending on what accounts for the fact that the eyewitness in PRISON YARD-(b) is 85% reliable, her testimony most likely either distributes the remaining 15% chance of error across an ad-hoc class of prisoners—the class of people who look like the actual attacker—or randomly. Either way, the imposition is a one-time risk that is not disproportionately concentrated. In that the evidence is testimony about Alex's behavior, it's plausibly somewhere on our scale of evidence that weakens, or at the very least does not strengthen, Alex's complaint against the risk of false conviction. So, we can expect that the stakes for justifying conviction are at least not raised by reliance on this evidence, and *might* be lowered.

The net effect is that there is a marked asymmetry between the standards for adequate justification to settle that p in cases like PRISON YARD-A than in cases like PRISON YARD-B, so even if the evidential probability of p in the former was slightly higher, it might fail to clear the demanding a-case threshold, while the latter easily clears the less demanding b-case threshold.

3 Justifying Exclusion

If all I've said so far is true, we can expect that purely actuarial inferences will rarely satisfy the RATIONAL SETTLING CONSTRAINT, and so we should expect asymmetries of the kind on display in the PRISON YARD case pair. But this does not yet justify *excluding* actuarial evidence from consideration at trials. For that, we will need to invoke moral reasons—but here they are appropriate. I take it that Tillers' complaint is not against the relevance of moral reasons *as such*, but rather against their inappropriately distorting our epistemic practices.

All rational agents must adopt practices to balance competing epistemic goals: *believe truths* and *avoid believing falsehoods*. But even if individuals are free to decide for themselves how to make these tradeoffs, institutional finders of fact are not. The purpose of the legal apparatus is to help fulfill the state's obligation to enforce and protect the rights of its members. Its governing policies must facilitate justice, not only narrowly in the outcome of a given case, but more broadly in each member's overall risk of suffering either of two bad outcomes: (i) a violation of their rights, or (ii) a false finding of guilt/liability. Decisions about the epistemic policies of

[18] The operation of the principle does not seem limited to explicit speech acts. What it tracks instead is something like *responsible performance of avoidable behaviors*, where something counts as avoidable if it is not costly, in some normatively laden sense, for the agent to avoid. In determining whether avoidance is costly, we should discount costs the agent is liable to bear, but count as costs any sacrifices of moral entitlements, and arbitrary or disproportionate restrictions on overall freedom, as well as expenditures of resources or effort.

this institutional agent–most obviously, identification of burdens and standards of proof–must be selected with reference to this goal. If the standards are too high, or too much evidence is excluded, then the expectation of escaping responsibility for one's crimes will be too high, and members will not be adequately secured against suffering violations. If the standards are too low, then they face too high a risk of being falsely convicted or found liable. Similarly, justice demands that the policies be crafted in a way that distributes the remaining risks fairly among the members. Schauer and Tillers' objection overlooks this point: when crafting the epistemic norms for the court, we cannot look only to maximize our accuracy in a given ruling, or even across a long stretch of rulings. We are also obligated to look at how our policies will affect the distribution of the errors we *do* make. If an evidence rule concentrates the risk of suffering false findings on a subgroup of the population, in a way insensitive to their choices, members of that group have a justice-based complaint against the rule proportionate to the severity and concentration of the risk.

This is because even group members who do not themselves suffer mistaken findings are harmed by disproportionate risk impositions of this kind.[19] The knowledge that they face disproportionately high risk of being found guilty or liable, if they face a trial, can be expected to have a number of harmful effects. First, simply being aware that one faces higher risks can take a psychological toll, affecting the agent's overall welfare and well-being. Second, when the costs associated with the risk are especially high—high fines, a conviction, detainment—disproportionate exposure can lead agents to engage in a wide array of costly behavior to protect themselves from the harm, or to constrain their actions in various cost-incurring ways in order to reduce their risk exposure. These effects of risk exposure are compounded by repeated, sustained, or patterned imposition. The more stable, visible, and identity-tracking a property G is, the more actuarial inferences based on G-membership will concentrate risk.

If G is a property that scores highly on these measures, an evidential policy that permits actuarial inferences based on G to count as evidence of guilt can be expected to influence G-members to act to minimize their risk of ever facing trial: avoiding contact with law enforcement or government agencies, avoiding public spaces where others might find their presence 'suspicious', etc. These behaviors are costly to the individuals as well as being bad over the long term for the polity.[20] So, we have particularly strong reason to bar reliance on actuarial inferences based on visible, identity tracking properties (race, gender, religion, orientation). If they were highly probative evidence, or if excluding them could be expected to significantly increase the incidence of errors, this might be a significant sacrifice requiring political deliberation. But, as we have just seen, actuarial inference *especially* based on these groups can be expected to generally be of minimal probative value.

Tillers is right, however, that this moral/political motivation for exclusion neither affects *all* actuarial inference, nor *only* actuarial inferences. This is one of the strengths of the explanation I have offered; the nuanced verdicts it yields about which forms of evidence should be eschewed explain some intuitive verdicts that are difficult for a simpler account (like STATISTICS DEFICIENCY) to accommodate. For instance, some find the PRISON

[19] There is some controversy over whether 'pure risks'–which do not eventuate in harms, and never come to A's attention–wrong A, but we can safely set this controversy aside. Our inquiry concerns whether we should adopt a policy of allowing legal convictions on the basis of statistical evidence, and evidential policies of this kind are public. It's therefore plausible that if we do adopt such a policy, the members of groups which thereby become more likely to be convicted for certain offenses may be aware of their increased risk. So, the sorts of risk impositions relevant to our discussion are 'impure', and can be counted as wrongs even if pure risks are not.

[20] Brayne (2017) details some of these behaviors among ethnic groups that are subject to disproportionate police scrutiny.

YARD-A case unproblematic, but would object if the statistical evidence invoked a racial or ethnic group rather than just the set of prisoners in the yard. The view I have sketched readily explains this pattern of judgments: if the group invoked by an actuarial inference is highly accidental or strongly choice-sensitive, the members will not have a weighty complaint against the risk imposed. So, to the extent that we take presence in the yard to be strongly choice-sensitive and accidental, we should be less bothered by basing inferences on this property. Conversely, and more interestingly, if individualized evidence like eyewitness testimony has a biased error distribution against a relatively stable group—e.g. makes false-positive identifications disproportionately often against black men—then members of that group have a justice-based complaint against relying on this form of evidence, proportionate in force to the bias of the error rate.[21]

The judicial system should be the just and fair way to collectively deliberate about and enforce public law. If the risks of suffering a miscarriage are not fairly distributed—either randomly or following choice in justifiable ways—then the system is not fulfilling its aims. What this highlights is that considerations centered on the particular individual on trial do not exhaust the moral difference between actuarial and individualized evidence. There is, then, an appropriate role for moral and political reasons in shaping our evidential policies, and, in these cases, following their prescription comes at no epistemic cost.

4 Wrapping up

The explanation that I have outlined abandons SIMPLE-PROBABILISM: it denies that statistical generalizations are straightforwardly equivalent to equally probabilifying individualized evidence in justificatory force. But it does not commit to STATISTICS-DEFICIENCY: it does not claim that there is a deep defect in actuarial inference as a class, and so escapes needing to provide a principled characterization of that class. Nor does it claim that evidence must be an appropriate basis for *believing* p, or yield knowledge that p, in order to justify issuing a legal verdict that p. Instead, I claim that the asymmetry arises from the fact that legal judgments are subject to the *rational settling constraint*, so whether evidence e is sufficient for *settling that* p depends on three factors: (i) how probable e makes p, (ii) how certain we are that future evidence won't significantly lower that probability (resilience), and (iii) what the expected costs are if we're mistaken whether p (practical adequacy). Population-level generalizations are a less resilient basis for credence than paradigmatically individualized information. They also have a harder time rendering agents practically adequate, because relying on reference-class based evidence raises the stakes by imposing a risk on other members of the class (which is weightier the more stable the class is, and the less random its membership).

My arguments suggest a framework for explaining the relevance of moral considerations to legal rules of evidence; it is not just wishful thinking, pretending the world were fairer and population-level facts didn't support certain probabilistic judgments. Rather, it emphasizes each individual's entitlement to a fair procedure and attends to how various evidential policies operate over time to shape our error propensities, taking these and not just the overall probability of errors to be relevant to agents' complaints. This is more nuanced, or at least more articulated at a couple of points, than STATISTICS DEFICIENCY. While it does acknowledge the importance of moral and political considerations in the framing of legal evidence rules, the explanation I offer makes clear how and when these factors are epistemically relevant.

[21] Thank Jessica Kieser and Katie Steele for this worry.

References

Allen, R. J., & Pardo, M. S. (2019). Clarifying relative plausibility: A rejoinder. *working paper*.

Anderson, C., & Hawthorne, J. (2019). Knowledge, practical adequacy, and stakes. In *Oxford studies in epistemology* (Vol. 6). Oxford: Oxford University Press.

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

Bolinger, R. J. (2018). The rational impermissibility of accepting (some) racial generalizations. *Synthese*, 1-17. https://doi.org/10.1007/s11229-018-1809-5.

Bratman, M. (1992). Practical reasoning and acceptance in a context. *Mind*, *101*(401), 1-15.

Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review*, *82*(5), 977-1008.

Brennan-Marquez, K. (2017). Plausible cause: Explanatory standards in the age of powerful machines. *Vanderbilt Law Review*, *70*.

Buchak, L. (2014). Belief, credence and norms. *Philosophical Studies*, *169*, 285-311.

Cohen, J. (1977). *The probable and the provable*. Oxford: Clarendon Press.

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

Duff, R. A. (1998). Dangerousness and citizenship. In A. Ashworth & M. Wasik (Eds.), *Fundamentals of Sentencing Theory: Essays in Honour of Andrew von Hirsch* (1st ed.). Oxford: Oxford University Press.

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). Legal burdens of proof and statistical evidence. In J. Chase & D. Coady (Eds.), *Routledge handbook of applied epistemology* (1st ed.). London: Routledge.

Haack, S. (2012). The embedded epistemologist: Dispatches from the legal front. *Ratio Juris*, *25*(2), 206-35.

Jackson, E. (2018). Belief, credence, and evidence. *Synthese*. doi: 10.1007/s11229-018-01965-1

Littlejohn, C. (2017). Truth, knowledge, and the standard of proof in criminal law. *Synthese*, 1-34. https://doi.org/10.1007/s11229-017-1608-4.

Locke, D. (2013). Practical certainty. *Philosophy and Phenomenological Research*, *90*(1), 72-95.

Moss, S. (2018). Moral encroachment. *Proceedings of the Aristotelian Society*, *118*(2), 177-205.

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

Pardo, M. S. (2019). The paradoxes of legal proof: A critical guide. *Boston University Law Review*, *99*.

Pritchard, D. (2017). Legal risk, legal evidence and the arithmetic of criminal justice. *Jurisprudence*, *9*(1), 108-119.

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

Risinger, D. M. (2004). Unsafe verdicts: The need for reformed standards for the trial and review of factual innocence claims. *Houston Law Review*, *21*, 1281.

Risinger, D. M. (2018). Leveraging surprise: What standards of proof imply that we want from jurors, and what we should say to them to get it. *Seton Hall Law Review*, *48*, 965-994.

Ross, J., & Schroeder, M. (2014). Belief, credence, and pragmatic encroachment. *Philosophy and Phenomenological Research*(2), 259-288.

Schauer, F. (2003). *Profiles, probabilities, and stereotypes*. Cambridge, MA: Belknap Press of Harvard University

Press.

Smith, M. (2017a). *Between probability and certainty: What justifies belief*. Oxford: Oxford University Press.

Smith, M. (2017b). When does evidence suffice for conviction? *Mind*, *127*(508), 1193-1218.

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

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), 33-49.

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