Episteme (2024), **21**, 1241–1264
doi:10.1017/epi.2023.27

episteme

ARTICLE

Algorithmic Decision-making, Statistical Evidence and the Rule of Law

Vincent Chiao

Faculty of Law, University of Toronto, Toronto, Canada
Email: vincent.chiao@utoronto.ca

(Received 9 August 2022; revised 24 January 2023; accepted 30 March 2023; First published online 29 May 2023)

Abstract

The rapidly increasing role of automation throughout the economy, culture and our personal lives has generated a large literature on the risks of algorithmic decision-making, particularly in high-stakes legal settings. Algorithmic tools are charged with bias, shrouded in secrecy, and frequently difficult to interpret. However, these criticisms have tended to focus on particular implementations, specific predictive techniques, and the idiosyncrasies of the American legal-regulatory regime. They do not address the more fundamental unease about the prospect that we might one day replace judges with algorithms, no matter how fair, transparent, and intelligible they become. The aim of this paper is to propose an account of the source of that unease, and to evaluate its plausibility. I trace foundational unease with algorithmic decision-making in the law to the powerful intuition that there is a basic moral and legal difference between showing that something is true of many people *just like you* and showing that it is true of *you*. Human judgment attends to the exception; automation insists on blindly applying the rule. I show how this intuitive thought is connected to both epistemological arguments about the value of statistical evidence, as well as to court-centered conceptions of the rule of law. Unease with algorithmic decision-making in the law thus draws on an intuitive principle that underpins a disparate range of views in legal philosophy. This suggests the principle is deeply ingrained. Nonetheless, I argue that the powerful intuition is not as decisive as it may seem, and indeed runs into significant epistemological and normative challenges. At an epistemological level, I show how concerns about statistical evidence's ability to track the truth can be resolved by adopting a probabilistic, rather than modal, conception of truth-tracking. At a normative level, commitment to highly individualized decision-making co-exists with equally ingrained and competing principles, such as consistent application of law. This suggests that the "rule of law" may not identify a discrete set of institutional arrangements, as proponents of a court-centric conception would have it, but rather a more loosely defined set of values that could potentially be operationalized in multiple ways, including through some level of algorithmic adjudication. Although the prospect of replacing judges with algorithms is indeed unsettling, it does not necessarily entail unreasonable verdicts or an attack on the rule of law.

Keywords: statistical evidence; rule of law; algorithmic decision-making; legal proof

Introduction

Law is a notoriously change-averse profession. Nonetheless, the rapidly increasing role of automation throughout the economy, culture and our personal lives seems unlikely to stop at the courthouse doors. Alarmed by this prospect, many have raised concerns about algorithmic decision-making: that it is biased, shrouded in secrecy, and difficult to interpret. These criticisms, while fair in particular contexts, are specific rather than general, implementational rather than foundational. They can largely be addressed by higher quality data, updated engineering priorities, and a better regulatory environment. They do not address the fundamental unease many feel about the bare prospect that we might one day replace judges with algorithms.

The objective of this paper is to investigate one route for vindicating the sense – one which I do not share, but which I suspect many do – that algorithmic decision-making is in some fundamental way inimical to our idea of what law is about, no matter how fair, transparent, or intelligible it becomes. A very similar unease, I argue, pervades much of the literature regarding statistical evidence in law. Perhaps the epistemological and normative reasons that many theorists give for doubting that "bare statistical evidence" can support a legal verdict will also generalize to algorithmic decision-making in law, as algorithmic decision-making often relies on complex forms of statistical inference. A theme in the philosophical literature on statistical evidence is that courts are meant to be venues for determining facts specific to the parties, resulting in highly tailored, individualized justice. Algorithmic decision-making seems to be the opposite: squashing everything unique about a case to fit a pre-determined statistical model, while discarding anything idiosyncratic. Human judgment attends to the exception; automation insists on blindly applying the rule.

Part 1 considers why algorithmic decision-making in law might be appealing, and why skepticism about statistical evidence might pose a fundamental challenge to it. In part 2, I lay out the intuitions that lie at the heart of statistical evidence skepticism and explain why those intuitions have a more limited reach than is sometimes appreciated. In parts 3 and 4, I consider and reject two prominent versions of statistical evidence skepticism. I shift gears in part 5 and consider an argument grounded in the rule of law. However, as I argue in part 6, the "anti-arbitrariness" argument turns out to face its own limitations.

The ultimate upshot is that a foundational unease with algorithmic decision-making in the law, while intuitive, turns out to be difficult to defend. Because the philosophical and legal literature in this area is extensive, I do not claim to provide a comprehensive argument for this conclusion. Nonetheless, the approaches I focus on are either recent and influential in legal epistemology, or grounded in widely-shared moral principles about the proper functioning of the legal system. Appreciating the limitations of those arguments suggests that consideration of the potential upsides to algorithmic decision-making in the law should not be prematurely foreclosed because of evocative, but also elusive, intuitions about individualized justice.

1. The Anodyne Thought and the Powerful Intuition

"Algorithmic decision-making" could potentially mean a lot of things, from completely automated decision processes with no human involvement at all, to those in which an algorithm merely provides salient information to a human decision-maker. For instance, a device that automatically identifies speeding vehicles could automatically send speeding tickets to registered owners, or a human officer could be required to

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

review the photo and reading to decide whether to issue a ticket. For my purposes, I shall use "algorithmic decision-making" to include tools that merely make recommendations to those that take final decisions as well as those that render final decisions, and shall not distinguish between legal contexts (civil, criminal, regulatory). In the context I am interested in – adjudication in court – fully algorithmic decision-making is still rare, with algorithmic predictions generally treated as inputs into human judgment (Engstrom *et al.* 2020; Coglianese and Ben Dor 2021).

The main selling point of algorithmic decision-making in the law is that it may improve the accuracy of unaided human decision-making. Accuracy is a central virtue of adjudication, in the sense that many other virtues in adjudication presuppose accurate decisions. For instance, algorithmic decision-making may turn out to be faster and cheaper than human decision-making, but those are only virtues if the decisions rendered are reasonably accurate.

The centrality of truth to adjudication suggests:

> *The Anodyne Thought*. We *always* have reason to value verdictive accuracy; that is, the reason to favor accurate verdicts is never defeated, even if it may in some contexts be outweighed by competing values.

A direct consequence of the Anodyne Thought is that we have reason to use an algorithmic tool if doing so enhances the overall accuracy of verdicts. The thought is anodyne because enhancing the accuracy of verdicts is desirable on any view of adjudication that accords a central place to truth, which is most of them (Ho 2008).[1]

The claim that accuracy is a *central* value of adjudication does not imply that accuracy is the *only* virtue of adjudication. Some features of the legal process answer to independent values, such as giving the parties a say; others, such as exclusionary rules, work at cross purposes to accuracy. The Anodyne Thought is agnostic as to whether truth is lexically prior to other adjudicative values (Enoch and Spectre 2019). It presupposes only that truth is an important adjudicative value, even if one that is ultimately balanced against other adjudicative values. If that is right, then categorical suspicion of algorithmic decision-making in law is misplaced, as its appropriateness would inevitably depend upon context-specific evaluation of the relative impact on accuracy and other adjudicative values. Depending on the context, it might not be worth sacrificing other values to secure a marginal gain in accuracy, but that does not show that there is no value in improving accuracy.

Undermining the Anodyne Thought thus requires something more than the true, but equally anodyne, observation that legal adjudication is sensitive to values other than accuracy. Perhaps something like:

> *The Powerful Intuition*. There is a very important distinction between showing that something is true of many people just like you and showing that it is true of *you*.

Those who are perturbed by the prospect that human judges might be replaced, in whole or in part, by algorithmic tools must find a more far-reaching justification for their skepticism than the now-familiar concerns centering on algorithmic bias, opacity, or uninterpretability. After all, human decision-making is itself frequently biased,

[1]"A distinctive feature of contemporary legal adjudication," Ho writes, "is its fact-orientation" (2008: 262).

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

opaque, and difficult to interpret. This leaves skepticism about algorithmic decision-making in law susceptible to the charge of applying a double standard, and also subject to the argument that gains in accuracy attributable to an algorithmic tool might in particular cases be so great as to outweigh those downsides.

The Powerful Intuition provides a more foundational basis for skepticism about algorithmic decision-making in law, especially with respect to contemporary machine learning algorithms. The algorithmic technologies behind currently state-of-the-art generative language and image processing algorithms (such as ChatGPT or DALL-E), are essentially highly trained pattern-matching algorithms – that is, they correlate a given input prompt with an expected output, accounting for a very large number of parameters and a very large base of training data (Shanahan 2022). They are designed to predict the desired output from a given prompt based on statistical correlations in their training data. A large language model will output "Neil Armstrong" when prompted with "The first person to walk on the moon was …" simply because those are the words most likely to follow upon the prompt in its training data (Shanahan 2022: 2). In other words, impressive as they are, these algorithms are not generating answers because they stand in the right causal relationship to the states of affairs that make those answers true. Rather, they are generating predictions about properties of their training data, namely which outputs are most strongly correlated with a given input.

Machine-learning algorithms are, for this reason, as much at odds with the Powerful Intuition as more traditional statistical models. While machine-learning algorithms are less overtly procrustean than standard statistical models, which extrapolate a predetermined pattern into novel contexts, they nonetheless render predictions about particulars by comparing those particulars to a mass of other, similar, contexts – this pattern of pixels is likely to be labelled a "dog," this constellation of factors is likely to result in "recidivism" – rather than making a judgment about a particular on its own merits.

Conflict with the Powerful Intuition on its own, however, does not show why algorithmic decision-making should be especially troubling in legal contexts. Even if relying on such tools obliterates the distinction between what is true in general and what is true in a particular case, why should that be of *special* concern in legal contexts?[2] The final step in the case for skepticism about algorithmic decision-making in the law draws upon a truism about the rule of law, which is that it requires a separation of powers between the executive, legislative and judicial branches (Heath 2020: Ch. 3). In contrast to the legislative, which is constitutionally forbidden from making law about particular individuals, the courts have as their central function resolving disputes by applying necessarily general rules to specific cases (US Constitution, Art. 1, ss.9–10). In considering how general legal rules apply to particular cases, a court must thus consider the unique circumstances and context associated with the dispute it is asked to adjudicate. This makes pattern-matching – extrapolating what is known to be true about other parties in other disputes to infer what is likely to be true about this party in this dispute – fundamentally at odds with judicial reasoning in a liberal democracy. Pattern-matching handles uncertainty by extrapolating a general rule, one that has worked well in the past, forward to the novel case. It does not consider the particular, and hence does not attend

[2]The contrast between what is true in general and what is true in a particular case is central to Schauer's work (Schauer 1991, 2006). Schauer's influence on my arguments in this paper will be evident to anyone familiar with those volumes.

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

to the exception – the *sui generis* case that does not fit the latent patterns in the training data. Until algorithmic tools can equal human judges in this respect, they cannot legitimately substitute for human judges, no matter how accurate, unbiased, transparent, or intelligible they become.

In short, the Powerful Intuition provides a more fundamental source for skepticism about algorithmic decision-making in law. It does not rest on contingent features about the relative bias, opacity or interpretability of an algorithm as compared with a human judge. Rather, it is grounded on a perceived conflict between the nature of how algorithmic prediction works, including in its most powerful contemporary machine learning forms, and a quite basic commitment in liberal democracies about the structure of government. Unlike concerns about bias, opacity or interpretability, the Powerful Intuition is plausibly incorporated in the rule of law. This makes it a more promising basis for rejecting the Anodyne Thought outright, not merely accommodating it by asking whether the gains from accuracy are worth the costs associated with using a particular algorithm.

This way of articulating the Powerful Intuition has the virtue that it connects skepticism about algorithmic decision-making in the law to more familiar skepticism about the use of "bare" statistical evidence in court. This suggests that skepticism about algorithmic decision-making in the law is not simply another instance of the law's conservatism, but rests on widely shared and familiar intuitions.

To illustrate, consider the following highly idealized cases:

> *Prison yard.* One hundred prisoners are exercising in the prison yard. Ninety-nine of them join a planned attack on a guard; one refrains. No evidence is available to distinguish the 99 guilty parties from the 1 innocent party. For any given prisoner present in the yard at that time, the probability is 0.99 that he participated in the attack. Jones, a prisoner present in the yard at the time of the assault, has been charged with participating in the assault. Should Jones be convicted? (Redmayne 2008)

> *Blue bus.* Jones is struck by a bus and injured. She cannot identify which bus hit her, and there are no witnesses. However, she has sued Blue Bus Co on the grounds that they operate 90% of the buses in the area. Blue Bus Co defends by pointing out that Red Bus Co owns the other 10% of buses, and it could well have been a red bus that struck Jones.

The statistical evidence in these cases is "bare" because it merely identifies a base rate of some property across a population, without discriminating between defendants who do and do not exemplify that property. Statistical evidence skeptics claim that courts ought not find for the plaintiff in either *prison yard* or *blue bus* and generalize this result to a more general principle that courts ought not decide cases based on statistical evidence alone. Instead, consistent with the Powerful Intuition, they argue that courts must rely on individualized evidence; for instance – an eyewitness who claims to have seen the defendant participate in the assault or to have seen a blue bus speeding away from the accident. Such evidence is not necessarily more reliable than statistical evidence; indeed, it may well be less reliable. However, unlike bare statistical evidence, individualized evidence speaks to what the defendant himself did, not what other people did.

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

The Anodyne Thought reflects the common-sense judgment that the point of a trial is to reach an accurate verdict, even if we might have some other expectations as well, such as respecting due process and giving parties an opportunity to be heard. This is meant to be a fairly minimal claim – just strong enough to motivate the thought that we have reason to take seriously the prospect of bringing algorithmic tools to bear in the courtroom if doing so enhances verdictive accuracy overall. I have suggested that more familiar concerns having to do with algorithmic bias, opacity, and interpretability are not sufficient to defeat the Anodyne Thought. The best case for principled resistance to algorithmic decision-making rests on the intuitive (if admittedly elusive) concept of individualized evidence. The Powerful Intuition, as I have labeled it, has three main features: first, it underwrites skepticism about algorithmic decision-making in law no matter how fair, transparent, and interpretable an algorithm becomes. Second, there is good reason to regard it as particularly compelling in legal contexts because of deeply engrained understanding of how the separation of powers figures in the rule of law. Finally, the Powerful Intuition explains the prevailing skepticism among philosophers and legal theorists about the use of statistical evidence in court, suggesting that there is a common thread linking skepticism about statistical evidence and skepticism about algorithmic decision-making in court.

For these reasons, an argument that draws on the Powerful Intuition to connect epistemological concerns with political principles is probably the most plausible way of defending a generalized skepticism about algorithmic decision-making in the law. My aim in the rest of this paper is to probe the limits of this argument, and to suggest where it may come up short. I start, in the next two sections, by critiquing its epistemological dimensions. I then turn to questioning its underlying conception of the rule of law.

2. Statistical Evidence Skepticism I: Clarifications and Limitations

While evocative, it is tricky to say exactly what the distinction between "statistical" and "individualized" evidence is supposed to be. Perhaps the most intuitive approach is by appeal to probabilistic generalization: "99 out of 100 prisoners participated in the assault" versus "I saw Jones hit the guard." But this clearly will not do. After all, eyewitness evidence also relies on probabilistic generalizations, namely generalizations about the accuracy of the witness's visual identification and recall capacities (Schauer 2006). If people were unable to reliably identify other people visually, remember that identification, and report it accurately later, then the entire practice of calling eyewitnesses to testify would hardly have any point. Similarly, a fingerprint found at the scene of the crime could be presented either as evidence connecting a specific individual to the crime or it could be presented statistically, for instance (as with DNA evidence) as the probability of finding another match from a defined population (Faigman *et al.* 2014: 438).

The same point emerges more clearly in the following pair of cases:

> *Statistical bail.* A risk assessment algorithm developed for purposes of bail determinations predicts risk (failure to appear for a court date, or commission of an offense while pending trial) by weighting a person's age, sex, and the nature of the crime for which the individual was arrested. A judge at a bail hearing defers to the risk assessment tool and orders an arrestee detained.

> *Judge bail.* At a bail hearing, a judge observes an arrestee's age, sex, and demeanor in court. The judge considers the nature of the crime for which that individual was

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

arrested, prior criminal record, lack of stable employment, and substance abuse problems. After considering these factors, the judge orders the arrestee detained.

There are many reasons one might give for distinguishing between *statistical bail* and *judge bail*. Probabilistic generalization is not among them. After all, the judge in *judge bail* is relying on probabilistic generalizations about the effect of substance abuse, employment, and demeanor just as much as the algorithm in *statistical bail*. The nature of these generalizations differs, but not in a way that helps statistical evidence skeptics. Perhaps the most salient difference is that the generalizations in *statistical bail* have been empirically validated, whereas those in *judge bail* depend on a single judge's personal experience and professional opinion.[3] And while it is true that the underlying data in *statistical bail* may be biased, that too does not distinguish the cases since a judge will see the same biases in the portfolio of arrestees brought before her. Consequently, the distinction between individualized and statistical evidence cannot be spelled out by appeal to decision making that does or does not rely on probabilistic generalizations. Probabilistic generalizations are ubiquitous.[4]

Another influential account posits that individualized evidence is *caused* by the defendant's actions – e.g., if he didn't commit the robbery, his DNA wouldn't be on the knife – whereas statistical evidence is not (Thomson 1986). However, statistical evidence often is causally related to the defendant's actions. Consider *prison yard*: if Jones participated in the assault, then the evidence would show that 99 out of 100 people in the yard are guilty, whereas if Jones had not participated in the assault the evidence would show that 98 out of 100 people are guilty (Gardiner 2018: 181; Smith 2018: 1204; Di Bello 2019: 1050).[5]

Rather than explore further attempts to move beyond an intuitive grasp of the contrast between individualized and statistical evidence, I propose taking a step back to consider the myriad circumstances in which courts unproblematically rely on statistical evidence. Doing so undercuts the suggestion that statistical evidence skeptics are merely reconstructing a clear and consistent judicial practice (Ross 2021: 10).[6] This does not show that statistical evidence skepticism is unwarranted, but it does suggest caution in generalizing from evocative, but potentially unrepresentative hypothetical cases, such as *blue bus* and *prison yard*.

First, it is important to distinguish between cases where a court is asked to evaluate a law, policy, or regulation, and those where a court is asked to apply a general rule to a particular instance (*Muller v Oregon* 1908). Statisical evidence is only even potentially controversial in the latter type of case, since the former type of case involves the anticipated effects of a law on a large group of people. Courts adjudicating constitutional challenges to a statute, for instance, will regularly consider statistical evidence about the statute's impact.

Second, sometimes disputed questions are inherently probabilistic. In these types of contexts "probable" is, as Haack puts it, "part of the content of the claim itself," rather

[3]Experience is positively correlated with accuracy, but the effect is limited (Spengler *et al.* 2009).

[4]"The exclusion of probabilistic evidence is impossible, because all evidence is probabilistic" (*United States v. Shonubi* 1995: 514).

[5]At a psychological level, there is suggestive evidence that reluctance to impose liability is not affected by whether the base rate statistic is caused by the defendant's actions (Wells 1992: 742).

[6]Schauer argues that *prison yard* presupposes an incorrect interpretation of how specified a legal complaint must be (Schauer 2021: 1). For a prominent Canadian case going the other way, see *R v. Thatcher* (1987).

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

than merely an acknowledgment of uncertainty: "S will probably do X" rather than "probably, S did X" (Haack 2014: 57–8). Bail is a case in point. In a bail determination, what must be proved is that the accused is at an elevated risk of fleeing the jurisdiction or committing a crime, not that he will. Or suppose a doctor is accused of wrongly reducing a person's chances of surviving an illness; in that case, statistical evidence about the probability of survival would seem plainly relevant. Or suppose a plaintiff argues that the defendant's product infringes on the plaintiff's trademark because it confuses consumers. The question is not whether any given consumer is confused, but whether consumers in the aggregate are. Courts commonly rely on statistical evidence about consumer beliefs in this type of case, and apparently have done so for fifty years (Monahan and Walker 2011: 75).

Third, statistical evidence is sometimes acceptable even in cases that involve deterministic phenomena, for instance in discrimination, mass tort, and anti-trust cases. In mass tort or employment discrimination cases, the plaintiffs are alleging that they each suffered a concrete harm (e.g., loss of wages or physical injury) at the hands of the defendant, but statistical evidence may be unavoidable if individualized proof is prohibitively expensive or time-consuming (Koehler 2002: 386–98; Monahan and Walker 2011: 75–6; Pardo 2019; *United States v Shonubi* 1995: 517). For instance, in the mass tort context, courts use statistical evidence as a means of achieving "the policy goal of compensating a tort victim when liability is clear but the identity of the wrongdoer is in question" (Koehler 2002: 400).

Fourth, courts do not categorically reject bare statistical evidence. DNA evidence is paradigmatic, for instance in establishing guilt or innocence or resolving a paternity suit. A key part of DNA evidence is the estimated frequency that an identified DNA profile will be found in a defined population (Ross 2021: 7). "Cold hit" DNA cases, where DNA evidence is used to identify (and potentially convict) a suspect, are similar to cases like *blue bus* and *prison yard* in that they rely on the improbability of finding a matching DNA profile by chance in the relevant population (Roth 2010). Cold hit DNA cases, like *blue bus* and some mass torts, are also cases where "liability is clear but the identity of the wrongdoer is in question." Outside the DNA context, there does not appear to be a clear practice, with courts sometimes allowing a case to proceed on base rate evidence and sometimes not.[7] A number of factors, including the quality of the statistical evidence, the defined nature of the reference class, the lack of alternative, individualized evidence, and a party asserting that an event occurred "by chance," play into whether a court will regard base rate evidence as relevant (Koehler 2002: 385–400).

Finally, sometimes it is the *refusal* to rely on bare statistical evidence that is controversial. Consider *McCleskey v Kemp* (1987). In that case, McCleskey argued that the state of Georgia's process for deciding whether to impose the death penalty was racially biased. McCleskey supported his claim with a statistical analysis that purported to show that people who kill white victims were far more likely to receive the death penalty than people who kill black victims. Rather than question the soundness of McCleskey's statistical claims, the Supreme Court held that even if his statistical analysis was sound, McCleskey had nevertheless failed to show that any individual official had discriminated against him. In other words, McCleskey's claim failed because he had only

[7]Compare *Kramer v. Weedhopper of Utah, Inc.* (1986) (reversing summary judgment for the defendant based on evidence that defendant supplied 90% of bolts to a manufacturer) with *Guenther v. Armstrong Rubber Company* (1969) (rejecting plaintiff's argument that defendant could be liable because it supplied 75–80% of the tires sold at a particular store).

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

statistical, rather than individualized, evidence of discrimination. Since he was unable to establish that he, personally, had been discriminated against, there was no reason to question the outcome on the particular facts of his case. McCleskey couldn't rule out the possibility that the prosecutor, judge, and jury simply regarded the facts of his case as sufficiently horrific to warrant death. Justice Powell, writing on behalf of the majority, emphasized that capital sentencing is a highly individualized process that requires "uniquely human judgments" to fully consider the "unique characteristics of a particular criminal defendant" (*McCleskey v Kemp* 1987: 311).[8]

A similar issue arose in *Dukes v Wal-Mart* (2011). This case involved an attempt to certify a class of 1.5 million female employees of Wal-Mart, who were alleging that Wal-Mart discriminated against them on grounds of sex. Lawyers for the plaintiffs tendered evidence showing statistically significant disparities in the percentage of women in management positions across Wal-Mart as a whole. However, Wal-Mart has a very decentralized approach to pay and advancement. Wal-Mart leaves personnel decisions to the discretion of local managers. The majority pointed out that the plaintiffs could not show that any of the 1.5 million women in the asserted class had been personally discriminated against on account of their sex (*Wal-Mart v Dukes* 2011: 356–8). The majority thus refused to allow the class action to proceed in part because the plaintiffs could only show disparate treatment across the class generally rather than individualized harm.

Both *Dukes* and *McCleskey* were 5–4 decisions.[9] The soundness of either judgment can reasonably be disputed. This suggests that even if there is something wrong with the use of bare statistical evidence in cases such as *blue bus* and *prison yard*, we cannot simply assume that skepticism in these hypothetical cases is automatically decisive in real life legal disputes.[10] Moreover, to the degree that it is problematic to rely on bare statistical evidence in court, that is only in cases involving the application of an established law to a deterministic phenomenon, in contexts where individualized litigation and fact-finding is not unreasonably costly or difficult and/or the quality of the statistical

[8]Justice Powell was, by his own admission, uncomfortable with statistical reasoning (Slobogin 2019: 130).

Some courts have taken a different approach to racial discrimination in sentencing. Canadian law requires sentencing judges to consider the discrimination Indigenous people have faced in Canada when sentencing Indigenous offenders, and in particular to consider non-custodial or generally milder sentences. Canadian courts have insisted that an Indigenous offender does not need to prove a "direct causal link" between his Indigenous status and the crime of conviction, and rather that courts may take notice of discrimination against Indigenous people in Canada as a general matter (*R v Ipeelee* 2012: para 83). The Supreme Court of Canada's emphasis on "background and systemic factors" which "may have" contributed to the crime, suggests that individualized evidence is not required, which in turn suggests that a statistical regularity – a high base rate of discrimination in a population – is sufficient. (*R v Ipeelee* 2012: para 82.) That said, lower courts faced with an Indigenous offender who grew up in fairly privileged circumstances or committed a crime that has no evident connection to his life experiences, have sometimes insisted on *some* kind of connection. (*R v. FHL* 2018: 83; *R v Bauer* 2013: 691.) It remains unclear what that means precisely, given the Supreme Court's rejection of a "direct causal" connection in favor of "background and systemic factors."

[9]*Dukes* split 5:4 on the commonality issue but was unanimous in holding that Federal Rule of Civil Procedure 23(b)(2) does not permit an action seeking individualized relief, such as money damages. (*Wal-Mart v Dukes* 2011: 360ff.)

[10]Blome-Tillman concedes that statistical evidence is appropriate in such circumstances but argues that this is because our intuitions of justice are not offended in such cases (Blome-Tillmann 2017: 279). A competing explanation is that people are willing to sacrifice truth when faced with statistical evidence that conflicts with sacred values (Tetlock *et al.* 2000).

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

evidence is doubtful. The willingness of courts to accept statistical evidence, including on grounds of practicality, arguably suggests that we should not read too much into intuitions elicited from hypothetical cases like *blue bus* and *prison yard*.

3. Statistical Evidence Skepticism II: Sensitivity

Holding the line against the Anodyne Thought requires a more robust case for statistical evidence skepticism. Probably the most sophisticated epistemological case is Enoch, Spectre and Fisher's analysis of statistical evidence in terms of sensitivity. On Enoch, Spectre and Fisher's construal, S's belief that p is sensitive if, "had it not been the case that p, S would (most probably) not have believed that p" (Enoch *et al.* 2012: 204).[11] Thus, in *prison yard*, the belief that Jones is guilty is not sensitive if founded on statistical evidence because that evidence would not change even if Jones were innocent. In contrast, believing that Jones is guilty because a witness claims to have seen him strike the guard is sensitive, because had Jones not struck the guard then (most probably) the witness would not have said that she saw him do so. Sensitivity provides a natural way of cashing out the Powerful Intuition because sensitivity requires evidence to track truth modally, not just probabilistically.

Because Enoch *et al.*'s overall argument about statistical evidence is dialectically complex, it is worth pausing to clarify what is – and what is not – at issue here. Enoch *et al.* argue that (a) statistical evidence is not sensitive; (b) knowledge requires sensitivity; however, (c) verdicts ought not be required to constitute knowledge (on pain of "knowledge fetishism"); nonetheless (d) statistical evidence should not be used in court because it generates perverse incentives for law-abidingness. My challenge is centered on (a): I defend a diagnostic, as opposed to modal, conception of sensitivity. I am agnostic on (b) and sympathetic to (c), but I take no firm position on either. This is because the Anodyne Thought is focused on truth, not knowledge.

This leaves Enoch *et al.*'s claim (d) that basing verdicts on statistical evidence weakens incentives to obey the law because someone contemplating whether to obey the law will realize that if he is included in the relevant reference class then his liability will depend upon the relevant base rate rather than his own conduct, assuming that he cannot control his membership in the reference class. For instance, a prisoner inside the prison yard (driver of a blue bus) will realize that he will be liable regardless of whether he attacks the guard (drives dangerously), weakening the law's deterrent effect (Enoch *et al.* 2012: 220ff.).[12]

Despite appearances, this claim is not in tension with my defense of the Anodyne Thought. Recall that the Anodyne Thought does not hold that the value of accuracy in verdicts is never outweighed; it holds only that it is never *defeated*, i.e., ceases to count as a reason. Taken at its strongest, Enoch, Spectre and Fisher's incentives-based argument does not defeat the reason we have to value verdictive accuracy. After all, Enoch, Spectre and Fisher are quite clear that we should not let "knowledge fetishism" get in the way of using statistical evidence in court if, somehow, we were able to blunt the effect of statistical evidence on incentives (Enoch *et al.* 2012: 221–2; Enoch and Fisher 2015: 584).

[11]Pardo describes sensitivity as a property of evidence rather than belief (Pardo 2018: 57). Since my analysis does not turn on this distinction, I will treat sensitivity as a property of both belief and evidence.

[12]Enoch, Spectre and Fisher note that the counterfactual reasoning that underpins the incentives argument parallels the role of sensitivity in underwriting knowledge claims (Enoch *et al.* 2012: 221).

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

Lest this seem a little too pat, it is worth considering how decisive the incentives argument actually is. To address this question, we need to answer two distinct empirical questions: first, what is the effect size – that is, how much does using statistical evidence in court weaken deterrence? And, second, how much does deterrence contribute to compliance? On the first question, there does not appear to be any reason to expect a uniform effect size, in large part because there is no reason to expect knowledge of legal processes to be evenly distributed. Repeat players, and sophisticated parties that are regularly advised by legal counsel, are likely to have different motivations than one-shotters and more naïve parties, who are less likely to attend to how courts use evidence. The degree to which people anticipate litigation is also relevant, as the use of statistical evidence will have little effect if people do not expect to wind up in court no matter what they do.

Secondly, fear of liability is clearly not the sole reason why people obey the law. Plausibly, people also obey the law for social and moral reasons, such as a belief in its legitimacy (Tyler 2006). Perceived legitimacy is likely to be especially important in contexts, such as criminal law, that engage highly salient norms of interpersonal morality. People may fear being punished for assaulting others, but the existence of robust and near-universal moral norms – both internalized and enforced socially – creates a strong, salient, and independent reason not to engage in such behavior. Insofar as social and moral reasons predominate in a given context, it becomes less important to sustain strong incentives based on fear of legal liability.

Taken together, the practical significance of a weakening of incentives due to the use of statistical evidence in court does not appear generally decisive, in the sense of so large as to dominate all other considerations. We might be willing to accept such weakening in some contexts but not in others. Recall that in some contexts, such as anti-discrimination and mass tort cases, courts have clearly decided in favor of allowing liability to turn on statistical evidence. Other cases, such as *Dukes* and *McCleskey*, are often controversial precisely because of disagreement about the relative importance of tailoring liability to individually identifiable instances of malfeasance as compared with providing relief for individuals who are highly likely to have been injured by the defendant's conduct.

In short, even without impugning the logic of the incentives argument, there are theoretical and empirical reasons to suspect that its practical significance is likely to vary by context, and in many cases to be significantly blunted by the existence of independent reasons favoring compliance.[13] This makes it difficult to say in the abstract whether the potential degradation of the deterrent signal due to the use of statistical evidence in court outweighs an incremental gain in verdictive accuracy. Consequently, the observation that accuracy is an undefeated value, even in contexts where it has to be weighed against competing values, is not a philosophical just-so story but rather a substantively plausible account of how we should deliberate about technological innovation in the law.

I now turn to considering the epistemological side of Enoch, Spectre and Fisher's argument, namely that statistical evidence lacks sensitivity. Enoch, Spectre and Fisher motivate their argument by appeal to the lottery paradoxes:

> *Statistical lottery*. On Monday, you buy a lottery ticket where the odds of winning are infinitesimally low. On Tuesday, the winning ticket is selected. Without

[13]For doubts about the logic of the incentives-based account, see Gardiner (2018: 185ff.).

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

checking to see what that winning ticket is, do you know that you did not win the lottery?

Individualized lottery. On Monday, you buy a lottery ticket where the odds of winning are very low. On Tuesday, the winning ticket is selected, and the winning number is printed in the newspaper. You look at the paper and see that your ticket was not selected. However, on rare occasions the newspaper misprints the winning lottery ticket number. The likelihood that on this occasion the newspaper has misprinted the winning number, and that in fact your ticket has won, is infinitesimally low. Do you know that you did not win the lottery? (Redmayne 2008: 297ff.; Enoch *et al.* 2012: 202ff.)

Many people report the intuition that you lack knowledge in *statistical lottery*, whereas you possess knowledge in *individualized lottery*. This is even though the likelihood of error is the same in both cases, namely infinitesimal. Sensitivity says: in *statistical lottery*, you would not learn you had won even if you did, whereas in *individualized lottery* if you won you would (almost certainly) learn about it. That is why your belief in *statistical lottery* does not count as knowledge and whereas your belief in *individualized lottery* does. By the same token, if we alter *prison yard* by adding that a prisoner testifies that he witnessed Jones participating in the assault, then we have sensitive evidence – an eyewitness would not (usually) say that Jones participated in the assault if he did not. In contrast, the statistical evidence on its own does not discriminate between the world in which Jones is guilty and the world in which he is innocent.

In explaining sensitivity's appeal, it is natural to appeal to a single counterfactual: what would happen if *this* case were the other way? Indeed, that is precisely the structure of hypothetical cases such as *blue bus*, *prison yard* and the lottery cases. It is also how Enoch and Fisher analyze sensitivity, namely in terms of the truth conditions of a single counterfactual. They explain that "a counterfactual is true, somewhat roughly if and only if its consequent holds in the *closest possible world* in which its antecedent is true" (Enoch and Fisher 2015: 609 n.182 (my emphasis)). On this analysis, a belief is sensitive if and only if we would not believe that p were we in the closest possible world in which p is false. This makes sensitivity into a discrete rather than continuous property. A belief is either sensitive or it is not.

However, thus understood, sensitivity has a very odd feature, which is that even if individualized evidence co-varies with ground truth in an individual case, there is no guarantee that it will do so more generally. Indeed, the paradox of statistical evidence is that following such evidence may actually do better at tracking truth across *most* cases, even if it fails to do so in a particular one. In other words, while sensitivity is analyzed in terms of a single counterfactual – what the evidence would show in the closest possible world in which it is false – the broader idea of evidence tracking the truth can instead be analyzed in terms of long run frequencies. This could be interpreted in terms of maximizing the total number of truthful beliefs (verdicts) in actual epistemic contexts, although more complex interpretations – assigning weights to different types of error, or prioritizing truth in the most salient counterfactual contexts – are available as well (Sosa 1999: 145–6; Pardo 2018: 60–1, 66).

I shall refer to a belief-forming mechanism's long run ability to track the truth as diagnosticity. I choose this term to draw attention to the battery of diagnostic tools commonly used to assess the reliability of a binary classifier, for instance, a medical test or an algorithmic risk assessment (Pines *et al.* 2013: Ch. 3; Fazel 2020). In a

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

diagnostic context, the accuracy of such a classifier is evaluated by considering how its false positive and false negative rate relate to each other. Two commonly used measures are "sensitivity" and "specificity" (Pines *et al.* 2013: 21). Note that sensitivity in this context differs from sensitivity in Enoch *et al.*'s sense: in the diagnostic context, it refers to the number of true positives a classifier generates divided by the sum of its true positives and false negatives, i.e., the actual number of positives. Specificity refers to the number of true negatives it generates divided by the sum of its true negatives and false positives, i.e., the actual number of negatives. Otherwise put, in a diagnostic sense, sensitivity measures a classifier's ability to pick out true positives, whereas specificity measures its ability to exclude true negatives. For instance, a COVID test's sensitivity tells you how good it is at detecting infected individuals, whereas its specificity tells you how good it is at ruling out non-infected individuals.

The same information can be used to compute other metrics. For instance, "positive predictive value" is a metric that evaluates the probability of p given a positive test for p. Conversely, "negative predictive value" evaluates the probability that not-p given a negative test for p.[14] A classifier with high positive predictive value is very likely to indicate p only if p, and a classifier with high negative predictive value is very likely to indicate not-p only if not-p. Moreover, the relationship between sensitivity and specificity can be graphically represented by plotting these values against each other. Plotting (1 – specificity), otherwise known as the false negative rate, along the y-axis and sensitivity (the true positive rate) along the x-axis yields what is known as the receiver operating curve, or ROC. A point on the ROC represents the probability that a randomly selected case will be accurately classified by the classifier (Pines *et al.* 2013: 23; Fazel 2020: 198).[15] A perfect classifier sits on the line (0, 1), meaning that it will correctly classify any randomly selected case, whereas a classifier that is no better than random will lie along the diagonal.[16]

In contrast to Enoch *et al.*'s analysis of sensitivity, diagnosticity is a continuous variable – something a classifier can have more or less of. This flows from the fact that diagnosticity is assessed by a classifier's performance across many cases, not just at a single counterfactual. For this reason, base rates matter to diagnosticity in a way that they do not for sensitivity in Enoch *et al.*'s sense. The predictive value of a test depends on the base rate of the property being tested for, even holding sensitivity and specificity constant. Suppose, for instance, that a test for X has 90% sensitivity and 80% specificity. In a population of 1000 where the base rate for X is 10% (i.e., 100 people are X), we will have 90 true positives, 180 false positives, 720 true negatives, and 10 false negatives. In this case, the positive predictive value is $90/(90+180) = 0.33$, and the negative predictive value is $720/(720+10) = 0.98$. Now suppose the base rate increases to 30%. In that case, the same instrument will yield 270 true positives, 140 false positives, 560 true negatives, and 30 false negatives, with a corresponding positive predictive value of 0.65 and negative predictive value of 0.95. In other words, as the base rate increases, so too does the likelihood that a positive result is a true positive, as well as the likelihood that a negative test result is just a fluke (Aitken *et al.* 2010: s.2.21; Pines *et al.* 2013: 24).

[14]Positive predictive value is (true positives)/(true positives + false positives). Negative predictive value is (true negatives)/(true negatives + false negatives) (Pines *et al.* 2013: 23).

[15]A related metric is the conviction ratio, or the "probability of convicting the innocent divided by the probability of convicting the guilty" (Hynes 2017: 19; Di Bello 2019).

[16]My thanks to Boris Babic for clarifications on the ROC.

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

Second, the base rate of a phenomenon is not necessarily set in stone. It may be responsive to how we define the population of interest, a point obscured by hypothetical cases such as *prison yard* that implicitly assume a population of interest. What seems troubling about *prison yard* is not so much that we convict an innocent person, since every fallible legal system will occasionally convict the innocent. Rather, it is that relying purely on the base rate makes it seem like we are making no effort to distinguish which one of the 100 prisoners in the yard was innocent. Put in diagnostic terms, convicting in *prison yard* would lead us – should all 100 prisoners be prosecuted – to 99 true positives, one false positive, and no true or false negatives. As a result, a classifier that says "believe someone is guilty if that person was present in the yard at time *t*" would lie at (1, 1) on the ROC – perfectly sensitive, but also entirely non-specific.

But why focus only on the people in the prison yard at the time of the assault? The setup of *prison yard* presupposes that we already know that people *not* in the yard had nothing to do with the assault, but surely whether that is so is an entirely contingent matter. Perhaps there is reason to suspect people who entered the prison yard prior to the assault, everyone elsewhere in the prison at the time, everyone within a 5 km radius of the prison, or even everyone in town. Suppose, in the latter case, that there are 9,900 people in town, meaning that the base rate is not 99 guilty people out of 100, but rather 99 out of 10,000, i.e., the people in town plus the people in the prison yard. In that case, our classifier, "believe someone is guilty if that person was present in the yard at time *t*" would yield 99 true positives, 1 false positive, 9900 true negatives, and 0 false negatives, with a corresponding sensitivity of 1 and specificity of 0.99. Expressed in terms of the ROC, applying our classifier to the entire population of the town identifies the point (0.01, 1), which is quite close to perfect classification. Framed this way, the classifier turns out to be extremely good at tracking the truth, both in terms of identifying those who are guilty and excluding those who are not.

Which framing is correct would seem to depend on our background assumptions and *ex ante* epistemic situation.[17] If we already know that the guilty parties are limited to those in the prison yard, then "believe someone is guilty if that person was present in the yard at time *t*" is non-specific, because it doesn't do anything further to rule out innocent parties. If we have reason to suspect the entire town is in on it, the classifier helps us rule out vast numbers of innocent parties, making it very specific indeed. The trouble with *prison yard* is that the case is designed to make it seem as if a classifier based on statistical evidence cannot be specific. But statistical evidence is only non-specific if we surreptitiously ignore everyone that the classifier *does* rule out.

[17]Di Bello (2019: 1050–1) also stresses the importance of background evidence, although he draws a different conclusion than I do. Di Bello suggests that because the probative value of statistical evidence depends on contextual evidence, it is wrong to convict based on statistical evidence alone as doing so leaves the risk of a wrongful conviction undetermined. While I agree that it is important to put statistical evidence into context, this is not what most statistical evidence skeptics are skeptical about, since they typically claim that what statistical evidence leaves out is individualized evidence of guilt, not information about a statistic's probative value in a particular context. Di Bello argues that the probative value of the statistical evidence is undetermined because he focuses on the likelihood ratio – how likely it is that we would obtain the statistic if the defendant were guilty versus innocent, accounting for background knowledge. To speak to that issue, a prosecutor needs to introduce evidence explaining why the statistic is both sensitive and specific: for instance, by explaining how it is we know that 99 out of the 100 people in the prison yard participated in the assault (sensitivity) and why we know that no one else did (specificity). That is important evidence, but it is not the individualized evidence that statistical evidence skeptics demand.

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

Obviously, if that move is legitimate, then every classifier ever is perfectly non-specific. The better conclusion is that the specificity of a classifier depends upon the identified base rate, and how we define the base rate depends upon the contingencies of our epistemic situation at the time. By the same token, a classifier cannot be rendered acceptable simply because it rules out large numbers of people who were never really in question anyway.

Finally, how does a focus on truth tracking over the long run, as assessed in diagnostic terms, handle individual cases? What makes *prison yard* challenging is that it focuses our attention on an individual case and asks whether we can exclude the possibility that *this* person might not be an exception to the general rule. We cannot. But this is not unique to statistical evidence. We also cannot exclude the possibility that an eyewitness is mistaken, a fingerprint was misread, or a confession falsely given. Rather, a focus on truth tracking over the long run asks: what pattern of results would we expect if we were to use the statistical evidence to prosecute everyone in the same position as the accused, would that pattern of results be within the appropriate margin of error, and how does it compare to the alternative? In *prison yard*, the two relevant options are one false conviction and 99 true convictions versus 100 false acquittals. Convicting based on the statistical evidence would maximize the number of truthful verdicts, but how important that is depends on the social cost of false acquittals relative to that of false convictions in the context in question, such as assault of a prison guard.[18]

Let me take stock. Enoch *et al.*'s appeal to sensitivity is attractive because it draws on the compelling principle that beliefs and evidence should track the truth, in the sense that we should notice (our beliefs should change) if the world were otherwise than it is. A sensitivity-based account is especially attractive in the context of skepticism about algorithmic decision-making in law, as it would allow skeptics to reject algorithmic tools based on statistical inference without having to deny the value of accurate verdicts. In contrast, I have argued that it is at least equally reasonable to focus on truth tracking overall rather than in particular cases. I illustrated this concern by reference to diagnosticity, as a suite of measures for assessing how well a classifier – such as a system of trials – tracks truth across many cases. Diagnosticity, however, suggests that it is entirely possible that attending to statistical evidence – and, by extension, replacing case-by-case human judgment with algorithmic judgment – could actually enhance the legal system's ability to track the truth. In fact, sensitivity, specificity and predictive value are regularly used to assess the performance of algorithmic prediction tools (Fazel 2020; Hester 2020). A significant part of tuning algorithmic risk assessment tools consists in trying to find the appropriate balance between sensitivity and specificity, with that relationship represented by the ROC (Hester 2020). Diagnosticity thus shows why a concern with truth tracking does not necessarily support skepticism about either statistical evidence or algorithmic decision-making in the law.

4. Statistical Evidence Skepticism III: Anti-arbitrariness

The Anodyne Thought says that because accuracy is a central virtue of adjudication, we have pro tanto reason to support any measure that improves the accuracy of

[18]This is true regardless of whether we initially suspected everyone in town or already knew that the perpetrators were limited to those present in the prison yard. That question bears on what we learn from the statistical evidence. The question of whether we should convict anyone, everyone, or no one depends on the degree and distribution of error in each of those cases.

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

adjudication. The Powerful Intuition says that there is an important distinction between showing that something is true of many people just like you and showing that it is true of you. The Powerful Intuition suggests that there is a categorical objection to using algorithmic tools that work by predicting what is (probably) true of a given individual on the basis of what is true of many other similarly situated people. Thus far, I have considered whether the appeal of the Powerful Intuition can be explained by reference to the concept of sensitivity. I distinguished between modal and diagnostic versions of sensitivity, and argued that even if the former supports the Powerful Intuition, the latter does not.

Perhaps, however, the Powerful Intuition can be defended by appealing to more overtly normative principles. For instance, Gardiner claims that

> If a court convicts a defendant erroneously, and the only response available is 'you win some you lose some' because the court relied on statistical evidence to satisfy the burden, the court has wronged the defendant. If a person is convicted, found liable, searched, or arrested without participating in the alleged activity, this error ought to arise from some abnormal feature. Plausibly justice – also public trust in the legal system – demands this (Gardiner 2018: 190).[19]

Yet this argument is mysterious on its face. What makes a wrongful conviction wrong is that *the defendant is innocent*. Someone who spends years in prison for a crime he didn't commit is not somehow wronged less if he is convicted because an eyewitness wrongly placed him at the scene and wronged more if he is convicted based on reliable statistical evidence. (Suppose he later files suit: do his damages depend upon whether the evidence used to convict him was individualized or statistical?) To my mind, a court that convicts while saying, in effect, "well *normally* eyewitnesses are right, even though often enough they aren't" reveals a greater disregard for the defendant's rights than a court that says, "our best (statistical) evidence leaves only a very small chance that you are innocent."

On Gardiner's view, a verdict based on statistical evidence is like concluding that your lottery ticket lost without even looking at it: probably right, but it feels like luck (Thomson 1986). Should we credit this feeling? We *do* occasionally describe a person's motivation in buying a lottery ticket as a desire to "try their luck." That said, at some level it is not really "bad luck" when a person who buys a lottery ticket with infinitesimal odds fails to win. At some level, "what did you expect?" is a completely reasonable response to someone who persists in complaining about her "bad luck" in not winning the lottery.[20]

Other normative arguments do not fare much better, as they often take crucial premises for granted – for instance, that a defendant has a "right" to an acquittal unless the prosecution can demonstrate that we know the defendant to be guilty (or civilly liable), and/or that there are no significant costs associated with acquitting the guilty

[19]Smith (2018: 1214) presents the argument in a more guarded fashion.

[20]Consider that the FBI's source attribution threshold for confirming an individual as the source of a DNA sample is one in 280 billion, or accounting for the total population of the United States, less than one tenth of one percent. FBI source attribution threshold (Roth 2010: 1138). Source odds calculated by multiplying random match probability (1 out of 280 billion) by total population of the United States (approximately 328 million.) Of course, in actual cases both the random match probability and the relevant reference class are likely to be quite a bit smaller.

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

(or denying recovery to valid legal claims) (Moss 2018: 215). Those premises are quite contentious, since they call into question the rationale for adjudicating legal claims to begin with. They are thus unlikely candidates for challenging the Anodyne Thought. Similarly, some philosophers have argued that beliefs about persons are subject to different moral norms than beliefs about non-persons, such as a special "rule of consideration" for the possibility that the person in question is an exception to a valid statistical generalization (Moss 2018: 221). However, those accounts falter in the face of ambiguity as to whether the object of a statistical generalization is a person (the driver of the blue bus) or not (the bus), and also presuppose moral principles (such as what it is required to treat someone as an individual) that are more controversial than the Powerful Intuition itself (Smartt 2022). This makes such accounts dialectically unsuited to unseat the common-sense claim that, whatever else we may want from trials, we always have reason to value accuracy in verdicts.[21]

Given these limitations, I now to turn to consider an argument that seeks to explain the power of the Powerful Intuition in a way that is specific to the legal context, namely by connecting that intuition to a familiar and widely shared understanding of the place of courts within the rule of law (Buchholtz 2020; Huq Forthcoming). That view implies that the province of the courts is to apply general rules to particular facts, a role that is undermined to the degree that courts resolve particular disputes by reference to evidence about what is generally true rather than what is true in a particular case. I shall refer to this as the argument from anti-arbitrariness. I start by sketching the argument from anti-arbitrariness in this section before turning to considering its limitations in the next section. Ultimately, I conclude that the argument from anti-arbitrariness also fails to conclusively reject the Anodyne Thought.

A recurring theme in the rule of law tradition is the contraposition of power and law. The rule of law tradition, especially in its Anglo-American variants, assigns courts the function of ensuring the lawfulness of executive action. For instance, Dicey famously extoled the English common law courts as a check on arbitrary government action, comparing them favorably to French administrative courts, which he suspected would merely provide legal justification for government overreach. Similarly, Hayek emphasized the stability and neutrality of common law courts as against the discretion of bureaucracies and officials to direct public policy, potentially in an idiosyncratic and unpredictable manner. In both versions, the distinction between law and power is lined up with the distinction between individual rights and the interests of the collective, with courts understood to be an essential means of vindicating the former as against the latter.[22]

Call this the "anti-arbitrariness" conception of adjudication, where a government action is arbitrary if it is not authorized by law.[23] Courts vindicate anti-arbitrariness by inspecting the lawfulness of an asserted exercise of public power in each and every case, something that the executive cannot do for itself. As Krygier (2019: 114) observes, hostility to arbitrary power is a "traditional reason" motivating the rule of law.

[21]Enoch and Spectre (2021) have recently argued that the problem of statistical evidence is inherently pluralistic, in the sense of resistant to unified philosophical analysis.

[22]As Beccaria (1995: 12–13) put it, while the sovereign may make general law, the sovereign "may not rule on whether an individual has violated the social pact." That requires a magistrate "to judge the truth of the matter."

[23]For a different account of arbitrariness, see Creel and Hellman (2022).

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

Anti-arbitrariness is most familiar in the context of high-profile political conflicts where courts are asked to step in to examine whether the government acts lawfully (*Black v Chrétien* 2001; *R (Miller) v Secretary of State for Exiting the European Union* 2017). But government power is asserted in myriad routine matters as well, and in such contexts the principle requires that courts be available to examine an assertion of government power to verify its lawfulness in each mundane instance. However, or so the argument might go, a court that relies on statistical evidence shirks its duty to examine the lawfulness of government action in every case, since it allows the government to proceed when the government has shown only that it is frequently entitled to act in many similar cases, but not that it is entitled to do so in *this* case. Yet if Jones did not participate in the assault, or Blue Bus Co did not actually own the bus involved in the accident, then the government has no lawful power to punish, or to enforce the state's demand for punishment or the plaintiff's demand for compensation.[24] Statistical evidence fails to ensure that government power is lawfully exercised in that very case, and instead authorizes coercive power simply because the government would be authorized to act in many other apparently similar cases. As Colyvan *et al.* (2001: 172) put it, "we require further evidence because the reference-class evidence is not specific to the individual in question."

A court that convicts on such evidence may not wrong the individual – after all, the evidence of guilt suggests the government has lawful power to proceed – but it nevertheless undermines the principle of anti-arbitrariness. Refracted through the rule of law, liability based on bare statistical evidence stands for the proposition that since the government would be authorized to act in a great many other cases that are very similar to this one, probably it is authorized to act in this case too. In other words, 'lots of people *just like Jones* broke the law, so it's safe to assume Jones did too.' One might regard this as walking back our commitment to the rule of law.[25] If anti-arbitrariness is essential to the rule of law, then perhaps we have the beginnings of an argument that vindicates the Powerful Intuition, and by extension, a foundational skepticism about algorithmic decision-making in the law.

5. Assessing Anti-arbitrariness

Rather than further unpack the argument from anti-arbitrariness, instead I briefly sketch three reasons to suspect that no version of it will be sufficient to unseat the Anodyne Thought.

First, even if it is valuable to ensure that government action is lawful, it is *also* valuable to ensure that adjudication tends to reach truthful outcomes. Recall that rejecting the Anodyne Thought requires more than noting that adjudication is responsive to a plurality of values. It requires giving one or more of those values strong, even lexical, priority over accuracy – a much more contentious proposition. While I do not claim that such a view is necessarily unreasonable – indeed, it is implied by views that assign either infinite disvalue to false conviction or zero disvalue to false acquittal – the conventional view is that while false positives are worse than false acquittals, they are

[24]Alex Stein (2005: 80–91) has emphasized the difference between probabilistic evidence of a general phenomenon rather than an individual case.

[25]Admittedly, the rule of law interpretation is a more natural fit for criminal than civil cases. One might finesse this by arguing that the rule of law should protect people not only from unlawful domination by government but also from unlawful domination by other people. See e.g. King (2012) and Krygier (2019: 131–2).

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

nonetheless worse along comparable scales, meaning that at some point some finite number of false acquittals will eventually outweigh a false conviction (Posner 2003: 67; Cheng and Pardo 2015: 202).[26] Indeed, a fallible legal system requires that this ratio be finite, since otherwise we would never impose liability on anyone for anything.[27]

Second, the principle that courts are required to treat like cases alike is also a constituent part of the rule of law, and this principle is in considerable tension with a strict view of anti-arbitrariness. The principle that courts must treat like cases alike requires courts to develop, even if only informally and implicitly, a general model that compares a given case to others along a series of dimensions, emphasizing some and disregarding others (Westen 1982; Cheng 2009). Treating like cases alike requires courts to generalize across cases rather than operate in an entirely particularistic manner; it requires courts to decide a given case in light of how other cases 'just like this one' were decided in the past (Schauer 1991). Uncomfortably, neither principle seems optional; both seem to be strongly associated with the rule of law.

Consider *Shonubi*, in which a drug courier was sentenced for importing heroin via JFK airport.[28] Applying the Federal Sentencing Guidelines' mandate to look at actual rather than charged conduct in setting sentence, the trial court considered evidence about the average amount of heroin trafficked by Nigerian drug couriers who flew through JFK. In applying this model, as Cheng has noted, the court implicitly discarded other features of the case – what about other traffickers who worked a day job as a toll booth collector? Or traffickers who are left-handed, under 6' tall, fly through LaGuardia, are Yoruba, or grew up in Kano (Colyvan *et al.* 2001: 172; Cheng 2009: 2082–3)? If we want to ensure that Shonubi is treated similarly to other cases, a court must develop some model of what features of Shonubi's case matter, and how much they matter. Reference class problems are ubiquitous, and hence "the mere fact that the subject of an inference is a member of multiple reference classes cannot block inference, for then all our inferences would be paralyzed" (Redmayne 2008: 287).

In this respect, *Shonubi* is no different from a run of the mill sentencing case. At a sentencing hearing, the parties will seek to draw the judge's attention to precedents that favor their side while distinguishing those emphasized by the other side. Part of the object of this argumentative practice is to honor the principle of treating like cases alike, as the lawyers propose competing theories of which cases the present one most resembles. Ultimately, the prevailing theory should show how the prior cases fit together in a sensible manner and hence why the past cases, when read as the prevailing theory suggests, are normative for *this* case. It is far from obvious that courts act contrary to the rule of law when they compare a given case to many others along a finite

[26]Cheng and Pardo (2015: 202) resist utilizing a loss function because it allegedly "prioritizes social welfare over accuracy," but their own approach avoids this charge only by focusing entirely on civil cases where the social costs of error are assumed to be equivalent, obviating the need to weight errors.

[27]Would it be a good idea to abolish the justice system? That is outside the scope of this paper and, more importantly, it is outside the scope of the anti-arbitrariness principle too. That principle says that courts get to examine the executive's actions for lawfulness, not that courts get to prevent the executive from enforcing the law altogether. That said, non-linear accounts of the error cost ratio, such as represented by a sigmoid function, may potentially better accommodate some of the intuitive deontological constraints in this area.

[28]For the original sentence imposed after trial, see *United States v Shonubi* (1992); the first appeal, *United States v Shonubi* (1993); the re-sentencing, *United States v Shonubi* (1995); the second appeal, *United States v Shonubi* (1997); the final sentencing, *United States v Shonubi* (1997).

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

number of dimensions. But if so, then it is not clear that there really is a *per se* objection to deciding one case by looking at what was decided in other cases.

Finally, interpreting the rule of law to require highly particularistic adjudication arguably overrates the role of courts. It is true that courts have historically held a vaunted place in the rule of law tradition. Recall Dworkin: "[t]he courts are the capitals of law's empire, and judges are its princes" (Dworkin 1986: 407). Of course, Dworkin was famously a champion of judges, but it is not just Dworkin; Raz and Waldron, for instance, have also insisted that the rule of law requires independent courts operating on accepted principles of natural justice or due process. Similarly, Dicey insisted that the rule of law not only requires that the government's arbitrary powers be constrained, but additionally that it is ordinary common law courts that do the constraining (Dicey 1885: 195–6). Yet, while it is plausible that what we conventionally call "the rule of law" requires, in part, that there be an independent check on the lawfulness of executive action, it is far less plausible that individualized fact-finding by courts is the only way we could possibly do this. In fact, courts do not invariably insist that they are the only ones who can possibly verify the facts that support liability, since they regularly defer to the evidence of experts (Faigman *et al.* 2014: 432–5).

Rule of law scholars have made this point in a more general way as well. For instance, Krygier has emphasized the "extremely variable ways" in which societies give effect to the rule of law (Krygier 2009: 47; 2019: 122). Similarly, Taekema and Huq have both noted that we ought not expect a principle as vague and contested as the "rule of law" to yield very specific institutional directives (Taekema 2021; Huq Forthcoming). Taekema points out that other legal traditions seem to be significantly less court-centric than the Anglo-American one. She argues that other conceptions of the rule of law give pride of place to legislatures as a means of reining in executive overreach, and that they are more concerned about challenges to the rule of law stemming from failures of effective government, such as corruption, than they are with defective court procedures (Taekema 2021: 41–4; see also Krygier 2009; Krygier 2019: 124; Hertogh 2015: 46–7). At the very least, we are owed evidence that things go horribly awry every time somebody tries some less court-centric way of checking the lawfulness of executive action.

More broadly, perhaps technological innovation will bring us to the point where, as Huq puts it, "[t]he social and human goods associated with the rule of law can be unbundled from the specific institutional forms with which they have come to be associated" (Huq Forthcoming: 11). And if better ways of providing those social and human goods become available, then surely it would be fetishistic to insist that we must not abandon the institutions we have become accustomed to, for no reason other than that we have become accustomed to them.

For a small-bore example, it does not seem beyond imagining that automated speed detection technology could become sufficiently reliable to warrant both automated ticketing and a strong presumption of lawfulness should an aggrieved driver challenge the ticket. Under such conditions, anti-arbitrariness would be vindicated by the quality of the technology rather than by the courts. It would be the generally accepted accuracy of the mechanism that protects people against the "fears, uncertainties and surprises" associated with the arbitrary exercise of power, not the prospect that a human judge will examine a human police officer's explanation for why she thought you were speeding (Krygier 2009: 65).

It is too early to say whether institutions other than courts can be equally effective at vindicating traditional rule of law principles, but it seems increasingly unreasonable to be confident that they cannot. Litigation is expensive, time-consuming, and

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

unpredictable; many people whose legal rights are in jeopardy cannot afford lawyers to advocate on their behalf; and it is in any case not obvious that everyone values elaborate hearings and trials to the same degree as those academics and lawyers who spend their time championing the rule of law. After all, interacting with judges and lawyers, potentially in a public courtroom, may be embarrassing: as Huq notes, for some "[i]t may be more respectful of autonomy and dignity to minimize the human interaction that comes with seeking a welfare benefit or contesting a misdemeanor" (Huq Forthcoming: 9). Consequently, while every society will have disputes about how to apply a general rule to particular cases, it seems unimaginative to insist that the only way to do this is through case-by-case contestation in court. Ultimately it is that very association that the rise of algorithmic decision-making in the law calls into question.

6. The Anodyne Thought Revisited

Many have a deep unease at the prospect that human lawyers and judges might one day be replaced by algorithms. This unease extends beyond familiar implementational concerns focused on bias, opacity, or interpretability. I have suggested that the underlying unease rests on the Powerful Intuition that showing that something is true of many people just like you is very different from showing that it is true of you. I have examined a variety of epistemological and normative arguments for parlaying this intuition into a case against algorithmic decision-making in the law.

I have argued that those arguments, while promising in various respects, are insufficient to reject the Anodyne Thought that we always have reason to value verdictive accuracy. Sensitivity is subject to a diagnostic interpretation based on long-run frequencies, and the rule of law resists unequivocal identification with particularistic decision-making in court. This leaves unease at the prospect of algorithmic decision-making unmoored from an epistemological or normative foundation.

Since I have neither sought to criticize the Powerful Intuition from first principles, nor to exhaustively canvas every possible argument that has been (or might be) given on its behalf, my conclusion – that the Anodyne Thought remains unrejected – is correspondingly qualified. That said, it is worth stressing again just how anodyne the Anodyne Thought is. Because accuracy is a central value of adjudication, we have reason to welcome algorithmic tools if they improve the overall accuracy of verdicts, even if – as is undoubtedly true – there are other values that matter in adjudication. While there are many important and urgent concerns pertaining to how algorithmic tools are implemented in high stakes legal contexts, the proper response to those concerns is to improve the algorithms, the data they are based on, and the regulatory environment in which they operate. As concerns about bias, opacity, and interpretability fade, our undefeated reason to prefer verdictive accuracy should lead to greater acceptance of algorithmic decision-making in the law.[29]

References

Aitken C., Roberts P. and Jackson G. (2010). *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses (Practitioner Guide 1)*. https://www.maths.ed.ac.uk/∼cgga/Guide-1-WEB.pdf.

[29]I am grateful to Boris Babic, David Enoch, Debbie Hellman and Martin Heslop for feedback on earlier drafts, participants at a Legal Theory workshop at the University of Virginia, and to April Lewtak and Aaqib Mahmood for research assistance. Work on this paper was supported by a SSHRC Insight Development Grant.

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

Beccaria C. (1995). *On Crimes and Punishments*. Cambridge: Cambridge University Press.
Blome-Tillmann M. (2017). 'More Likely Than Not': Knowledge First and the Role of Bare Statistical Evidence in Courts of Law.' In A. Carter, E.C. Gordon and B. Jarvis (eds), *Knowledge First: Approaches in Epistemology and Mind*, pp. 278–92. Oxford: Oxford University Press.
Buchholtz G. (2020). 'Artificial Intelligence and Legal Tech: Challenges to the Rule of Law.' In T. Wischmeyer and T. Rademacher (eds), *Regulating Artificial Intelligence*, pp. 175–89. Cham: Springer.
Cheng E.K. (2009). 'A Practical Solution to the Reference Class Problem.' *Columbia Law Review* **109**(8), 2081–105.
Cheng E.K. and Pardo M.S. (2015). 'Accuracy, Optimality and the Preponderance Standard.' *Law, Probability and Risk* **14**(3), 193–212.
Coglianese C. and Ben Dor L.M. (2021). 'AI in Adjudication and Administration: A Status Report on Governmental Use of Algorithmic Tools in the United States.' *Brooklyn Law Review* **86**(3), 791–838.
Colyvan M., Regan H.M. and Ferson S. (2001). 'Is it a Crime to Belong to a Reference Class?' *Journal of Political Philosophy* **9**(2), 168–81.
Creel, K.A. and Hellman D. (2022). 'The Algorithmic Leviathan: Arbitrariness, Fairness and Algorithmic Decision-Making Systems.' *Canadian Journal of Philosophy* **52**(1), 26–43.
Di Bello M. (2019). 'Trial by Statistics: Is a High Probability of Guilt Enough to Convict?' *Mind* **128**(512), 1045–84.
Dicey A.V. (1885). *Introduction to the Study of the Law of the Constitution*. London: Macmillan and Co.
Dworkin R. (1986). *Law's Empire*. Cambridge, MA: Harvard University Press.
Engstrom D.F., Ho D.E., Sharkey C.M. and Cuéllar M.-F. (2020). *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. https://law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.
Enoch D. and Fisher T. (2015). "Sense and 'Sensitivity': Epistemic and Instrumental Approaches to Statistical Evidence.' *Stanford Law Review* **67**, 557–611.
Enoch D. and Spectre L. (2019). 'Sensitivity, Safety, and the Law: A Reply to Pardo.' *Legal Theory* **25**(3), 178–99.
Enoch D. and Spectre L. (2021). 'Statistical Resentment.' *Synthese* **199**(3–4), 5687–718.
Enoch D., Spectre L. and Fisher T. (2012). 'Statistical Evidence, Sensitivity, and the Legal Value of Knowledge.' *Philosophy and Public Affairs* **40**(3), 197–224.
Faigman D.L., Monahan J. and Slobogin C. (2014). 'Group to Individual (G2i) Inference in Scientific Expert Testimony.' *University of Chicago Law Review* **81**(2), 417–80.
Fazel S. (2020). 'The Scientific Validity of Current Approaches to Violence and Criminal Risk Assessment.' In J.W. de Keijser, J.V. Roberts and J. Ryberg (eds), *Predictive Sentencing: Normative and Empirical Perspectives*, pp. 197–212. Oxford: Oxford University Press.
Gardiner G. (2018). 'Legal Burdens of Proof and Statistical Evidence.' In D. Coady and J. Chase (eds), *The Routledge Handbook of Applied Epistemology*, pp. 179–95. London: Routledge.
Haack S. (2014). *Evidence Matters: Science, Proof and Truth in the Law*. Cambridge: Cambridge University Press.
Heath J. (2020). *The Machinery of Government*. Oxford: Oxford University Press.
Hertogh M. (2015). 'Your Rule of Law is not Mine: Rethinking Empirical Approaches to EU Rule of Law Promotion.' *Asia Europe Journal* **14**, 43–59.
Hester, R. (2020). 'Risk Assessment at Sentencing: The Pennsylvania Experience.' In J.W. de Keijser, J.V. Roberts and J. Ryberg (eds), *Predictive Sentencing: Normative and Empirical Perspectives*, pp. 213–38. Oxford: Oxford University Press.
Ho H.L. (2008). *A Philosophy of Evidence Law*. Oxford: Oxford University Press.
Huq A. (Forthcoming). 'Artificial Intelligence and the Rule of Law.' In M. Sevel (ed.), *The Routledge Handbook of the Rule of Law*. London: Routledge.
Hynes R.M. (2017). 'From Blackstone's Ratio to the Conviction Ratio.' Virginia Public Law and Legal Theory Research Paper No. 17, 1–25. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2722797.
King J. (2012). *Judging Social Rights*. Cambridge: Cambridge University Press.
Koehler J.J. (2002). 'When do Courts Think Base Rate Statistics are Relevant?' *Jurimetrics* **42**(4), 373–402.
Krygier M. (2009). 'The Rule of Law: Legality, Teleology, Sociology.' In G. Palombella and N. Walker (eds), *Relocating the Rule of Law*, pp. 45–70. Oxford: Hart Publishing.

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

Krygier M. (2019). 'The Rule of Law and State Legitimacy.' In W. Sadurski, M. Sevel and K. Walton (eds), *Legitimacy: The State and Beyond*, pp. 106–36. Oxford: Oxford University Press.
Monahan J. and Walker L. (2011). 'Twenty-Five Years of Social Science in Law.' *Law and Human Behavior* **35**(1), 72–82.
Moss S. (2018). *Probabilistic Knowledge*. Oxford: Oxford University Press.
Pardo M.S. (2018). 'Safety vs. Sensitivity: Possible Worlds and the Law of Evidence.' *Legal Theory* **24**(1), 50–75.
Pardo M.S. (2019). 'The Paradoxes of Legal Proof: A Critical Guide.' *Boston University Law Review* **99**, 233–90.
Pines J.M., Carpenter C.R., Raja A.S. and Schuur J.D. (2013). *Evidence-Based Emergency Care: Diagnostic Testing and Clinical Decision Rules*, 2nd edition. Chichester: Wiley.
Posner R.A. (2003). *Law, Pragmatism, and Democracy*. Cambridge, MA: Harvard University Press.
Redmayne M. (2008). 'Exploring the Proof Paradoxes.' *Legal Theory* **14**, 281–309.
Ross L. (2021). 'Rehabilitating Statistical Evidence.' *Philosophy and Phenomenological Research* **102**(1), 3–23.
Roth A.L. (2010). 'Safety in Numbers? Deciding when DNA Evidence Alone is Enough to Convict.' *New York University Law Review* **85**, 1130–85.
Schauer F. (1991). *Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*. Oxford: Oxford University Press.
Schauer F. (2006). *Profiles, Probabilities and Stereotypes*. Cambridge, MA: Harvard University Press.
Schauer F. (2021). 'Statistical Evidence and the Problem of Specification.' *Episteme* **19**, 1–10.
Shanahan M. (2022). 'Talking About Large Language Models.' arXiv preprint. https://doi.org/10.48550/arXiv.2212.03551.
Slobogin C. (2019). 'The Use of Statistics in Criminal Cases: An Introduction.' *Behavioral Sciences and Law* **37**(2), 127–32.
Smartt T. (2022). 'Reconsidering the Rule of Consideration: Probabilistic Knowledge and Legal Proof.' *Episteme* **19**, 303–18.
Smith M. (2018). 'When Does Evidence Suffice for Conviction.' *Mind* **127**(508), 1193–218.
Sosa E. (1999). 'How to Defeat Opposition to Moore.' *Philosophical Perspectives* **13**, 141–53.
Spengler P.M., Messick S.J. and Schum D.A. (2009). 'The Meta-Analysis of Clinical Judgment Project: Effects of Experience on Judgment Accuracy.' *The Counseling Psychologist* **37**(3), 350–99.
Stein A. (2005). *Foundations of Evidence Law*. Oxford: Oxford University Press.
Taekema S. (2021). 'Methodologies of the Rule of Law Research: Why Legal Philosophy Needs Empirical and Doctrinal Scholarship.' *Law and Philosophy* **40**, 33–60.
Tetlock P.E., Kristel O., Elson B. and Green M.C. (2000). 'The Psychology of the Unthinkable: Taboo Trade-Offs, Forbidden Base Rates, and Heretical Counterfactuals.' *Journal of Personality and Social Psychology* **78**(5), 853–70.
Thomson J.J. (1986). 'Liability and Individualized Evidence.' *Law and Contemporary Problems* **49**(3), 199–220.
Tyler T. (2006). *Why People Obey the Law*. Princeton, NJ: Princeton University Press.
Wells G.L. (1992). 'Naked Statistical Evidence of Liability: Is Subjective Probability Enough?' *Journal of Personality and Social Psychology* **62**(5), 739–52.
Westen P. (1982). 'The Empty Idea of Equality.' *Harvard Law Review* **95**(3), 537–96.

US Constitution

R v. Bauer, [2013] ONCA 691.
Black v. Chrétien, 54 O.R. (3d) 215, [2001] O.J. No. 1853 (Ontario).
R v. FHL, [2018] ONCA 83.
Guenther v. Armstrong Rubber Company, 406 F.2d. 1315 (3d Cir 1969).
Kramer v. Weedhopper of Utah, Inc., 490 N.E.2d 104 (Ill. App. Ct. 1986).
R v. Ipeelee, [2012] SCC 13.
McCleskey v. Kemp, 481 U.S. 279 (1987).
R (*Miller*) *v. Secretary of State for Exiting the European Union*, [2017] UKSC 5.
Muller v. Oregon, 208 U.S. 412 (1908).

https://doi.org/10.1017/epi.2023.27 Published online by Cambridge University Press

United States v. Shonubi, 802 F. Supp. 859 (E.D.N.Y. 1992).
United States v. Shonubi, 998 F. 2d 84 (2d Cir. 1993).
United States v. Shonubi, 895 F. Supp. 460 (E.D.N.Y. 1995).
United States v. Shonubi, 103 F. 3d 1085 (2d Cir. 1997).
United States v. Shonubi, 962 F. Supp. 370 (E.D.N.Y. 1997).
R v. Thatcher, 1 SCR 652 (1987).
Wal-Mart Stores Inc. v. Dukes, 564 U.S. 338 (2011).

Vincent Chiao, B.A. (University of Virginia), Ph.D. (Northwestern), J.D. (Harvard), researches and teaches primarily in the area of criminal law and criminal justice, with a particular interest in the philosophical examination of its doctrine and institutions. He is the author of *Criminal Law in the Age of the Administrative State* (Oxford University Press, 2018).

Cite this article: Chiao V (2024). Algorithmic Decision-making, Statistical Evidence and the Rule of Law. *Episteme* **21**, 1241–1264. https://doi.org/10.1017/epi.2023.27