DOI: 10.1111/phis.12208

ORIGINAL ARTICLE

Philosophical Issues
A Supplement to NOUS

Standards and values

Matthew Kotzen

Department of Philosophy, University of North Carolina at Chapel Hill

Correspondence
Matthew Kotzen, Department of Philosophy, University of North Carolina at Chapel Hill.
Email: kotzen@email.unc.edu

1 | INTRODUCTION

Which values are at stake when we select a standard of proof to govern a particular legal determination, and how should these values inform the selection of a standard of proof? The matter is not as straightforward as has sometimes been suggested. There is little doubt that one central set of values at stake in the selection of a standard of proof includes those related to *error avoidance and allocation*: it is important both to minimize overall error in a system of adjudication and to allocate the errors that do inevitably arise in a manner reflective of their relative seriousness, and various procedural and evidential choices can have a direct impact on both of these goals. But there are a variety of other important values at stake in setting a standard of proof—involving the role that the standard plays in a multi-step system of adjudication, the incentives and disincentives that the standard provides for both "primary" and "secondary" conduct, the impact of the standard of proof on the expenditure of judicial resources and on the game-theoretic balance between parties, and what I will call the "epistemic rights" of the parties—that must also be accounted for. The goal of this paper is to explore the connection between these values and the choice of a standard of proof, and to provide a framework for the analysis of that choice.

2 | PRELIMINARIES

A crucial assumption that I will be making throughout this paper is that there is a coherent notion of a legal "error"—i.e., that there is some factual matter for the factfinder in a legal adjudication to get either right or wrong. This assumption can be questioned in a variety of ways, including on the grounds that some legal determinations are best understood as policy decisions (for example, about how to apply a rule or standard to a novel situation), rather than as adjudications of facts; though a policy can be better or worse than another policy, it is not clear that it makes sense to talk about a policy decision being erroneous in the relevant sense. However, I will assume here that, at least in the vast majority of cases, there is a fact of the matter about whether the defendant "really did it," and that an adjudication is either accurate or not depending on whether it matches the facts.[1] Indeed, the U.S. Supreme Court has repeatedly described the discovery of truth as "a fundamental goal of our legal system"[2] and as "the central purpose of a criminal trial."[3]

I will use the term "responsibility" to refer to the status of a particular (civil or criminal) defendant having *actually* acted illegally, and I will use the term "liability" to refer to the legal *determination* that they acted illegally. Then, we have a "false negative" when a responsible defendant is found to be non-liable, and we have a "false positive" when a non-responsible defendant is found to be liable. The false positive likelihood (FPL) is the conditional probability, assuming that a defendant is non-responsible, that they will be found liable. The false negative likelihood (FNL) is the conditional probability, assuming that a defendant is responsible, that they will be found non-liable.

A further assumption, both in previous approaches to the topic and in the one that I am taking here, is that legal factfinders have at least *some* degree of accuracy in their verdicts that is statistically higher than what would be produced by a chance mechanism (such as flipping a coin) to determine legal liability. Of course, we need not assume that factfinders are perfectly reliable, or anything close to it; indeed, the entire approach to legal factfinding that is under consideration here is *premised* on the assumption that factfinders are *not* perfectly reliable and hence that there will be errors which need to be appropriately distributed. The standard assumption here is that the overall strength of the evidence against both responsible and non-responsible defendants is (approximately) normally distributed, and that the mean of this (approximate) normal distribution is higher for responsible defendants than it is for non-responsible defendants; in other words, statistically, responsible defendants will tend to "look more responsible" to factfinders than non-responsible defendants will.[4]

A related (and fairly standard) assumption that I make here is that a standard of proof is a one-dimensional matter, corresponding to the linear "strength" of evidence that is required for a finding of liability; correspondingly, a factfinder is committed to finding liability when it determines that the net strength of the available evidence meets or exceeds the relevant threshold.[5] On this picture, the three most common standards of proof—preponderance of the evidence, clear and convincing evidence, and proof beyond a reasonable doubt (BRD)—can be placed (in that order) on a linearly increasing scale corresponding to the demandingness of each threshold. Since, by assumption, the mean of the strength-of-evidence likelihood distribution is higher for responsible individuals than it is for non-responsible individuals (and the shapes of these distributions is the same), *any* (non-extreme) threshold will tend to lead to a higher proportion of individuals being found liable among the responsible group than among the non-responsible group, for a fixed population of individuals.

Finally, though many of the same kinds of issues arise in the allocation of *burdens of proof* as arise in the selection of *standards of proof*, there are significant complexities involving burdens that require separate treatment.[6] For this reason, I will not address burdens specifically in this paper, though I do think that much of what I say here about standards can be extended to burdens. Indeed, my expectation is that a lot of what I say here can be carried over to other evidential and procedural rules, including various components of the Federal Rules of Civil Procedure, the Federal Rules of Criminal Procedure, and the Federal Rules of Evidence, though I cannot attempt a complete treatment here.

3 | ACCURACY AND TRADEOFFS

Once we have on board a notion of accurate and inaccurate verdicts, it is natural to understand the trial process as a sort of testing procedure analogous to testing procedures in science and medicine—here, involving a "measurement" by the factfinder of the net strength of the evidence

against the defendant and a "testing protocol" consisting of a comparison of that net strength to a particular threshold: the applicable standard of proof. Moreover, in designing such a testing procedure, it is natural to think both about how to increase the overall accuracy of the procedure, and about how to weigh against each other the various kinds of errors that can result from the procedure.

The idea that a standard of proof—especially in the context of criminal law—centrally involves a tradeoff between different types of legal error has a long history in Western thought. Its most famous formulation is Blackstone's pronouncement that "[i]t is better that ten guilty persons escape than that one innocent suffer,"[7] and hence discussions of the idea often refer to "Blackstone's ratio." However, the idea (albeit with somewhat different ratios) had historical precursors in the work of Aristotle,[8] Hale,[9] Fortescue,[10] and Maimonides,[11] among others. Justice Harlan, concurring in *In re Winship*, presented a more general formulation:

> Because the standard of proof affects the comparative frequency of these two types of erroneous outcomes, the choice of the standard to be applied in a particular kind of litigation should, in a rational world, reflect as assessment of the comparative social disutility of each.... In this context, I view the requirement of proof beyond a reasonable doubt in a criminal case as bottomed on a fundamental value determination of our society that it is far worse to convict an innocent man than to let a guilty man go free.[12]

In some situations and for some kinds of tests, it is possible to reduce *both* the FPL *and* the FNL for a particular type of test. Consider, for example, a medical test that is designed to diagnose disease *D*; a false positive is constituted by a positive result associated with a patient who does not have *D*, and a false negative is constituted by a negative result associated with a patient who does have *D*. Suppose that a particular testing procedure, though fairly reliable, is associated with a significant risk of contamination of any particular blood sample by another patient's blood, leading to a moderately high FPL and FNL. In such a case, improving the testing procedure by reducing the risk of contamination might improve (i.e., lower) both the FPL *and* the FNL, and thus represent a pure boon in terms of accuracy. Therefore, there is no reason *in general* why there must be a tradeoff between improvements in the FPL and the FNL for a testing procedure.

Similarly, there are some features of legal adjudicatory systems that plausibly reduce both the FPL and the FNL. For example, it is plausible that the requirement of sworn testimony—i.e., the requirement that testimony offered in court be under oath and under penalty of perjury—can be justified by its tendency to decrease both the FPL and the FNL. When witnesses are subject to penalty of perjury for lying, they will presumably lie less often. And although perjured testimony can, in principle, be used to support a meritorious party, it is reasonable to expect that non-meritorious parties—e.g., civil plaintiffs with a fraudulent tort claim, or criminal defendants trying to fabricate an alibi—are more likely to resort to the deliberate use of materially false testimony. Thus, by subjecting witnesses to the penalty of perjury, we disincentivize deliberately false material testimony, which differentially impacts the meritorious and the non-meritorious party, giving an advantage to the meritorious party. And since this is so regardless of which party is the meritorious party, it is natural to expect the requirement of sworn testimony to reduce both the likelihood that (meritorious) non-responsible defendants will be found liable, and also the likelihood that (non-meritorious) responsible defendants will be found non-liable, and hence to lower both the FPL and the FNL. Indeed, many basic procedural requirements, such as the requirement of sworn testimony, are almost entirely uncontroversial, in large part because they do not involve

any accuracy-related *tradeoff* or *balancing* among competing goods; they too are "pure boons" in terms of accuracy.[13]

By contrast, other features of adjudicatory systems—or, indeed, of *any* testing procedure—do involve deep and difficult tradeoffs between different dimensions of accuracy. In particular, when the test at issue consists of a measurement of a single quantity and a determination of whether that quantity meets or exceeds a particular threshold, then it will be impossible, through the adjustment of that threshold alone, to achieve decreases in both the FPL and FNL. Suppose, for example, that the medical test for disease *D* is simply a matter of testing the patient's (stipulated to be uncontaminated) blood to find the concentration of substance *S*, and determining whether that concentration is at or above a particular threshold. Then, no adjustment of the threshold up or down will ever achieve a decrease (or, for that matter, an increase) in both FPL and FNL. By moving the entire threshold up for everyone who takes the test, the effect will be to reduce the FPL and to increase the FNL; similarly, by moving the entire threshold down for everyone who takes the test, the effect will be to reduce the FNL and to increase the FPL.

It is similarly impossible in the legal context, through the adjustment of the standard of the proof alone, to achieve a decrease in either FPL or FNL without an increase in the other. Recall that I am helping myself to the assumption that standards of proof can be linearly ordered from lowest to highest. On this assumption, the likelihood of legal liability, both for responsible and non-responsible individuals, increases as the standard of proof is "relaxed" downwards, and decreases as the standard of proof is "tightened" upwards.[14]

4 | MULTI-STAGE ANALYSIS

Another important and relevant feature of many systems of adjudication—which often fails to play a proper role in analyses of errors in adjudication—is that they are conducted in multiple steps, rather than in a single step consisting only of the factfinder applying the standard of proof to the evidence that has been admitted at trial.[15]

In the criminal context, both an arrest and a formal charge require determinations that there is *probable cause* to support the charge; though jurisdictions vary in their definitions of probable cause, one common definition of probable cause is "evidence sufficient to induce a person of ordinary prudence and caution conscientiously to entertain a reasonable belief that the defendant committed the crime charged."[16] Moreover, there is universal agreement that the probable cause standard is less demanding than the BRD standard, and near-universal agreement that the standard is also less demanding than the preponderance standard.[17] In the U.S. federal system, this standard can be applied at a number of different junctures in the progress of a criminal case. If a warrant is issued before the defendant is arrested, then the magistrate issuing the arrest warrant is required to make a probable cause determination; if no warrant is issued before arrest, then probable cause is determined by a magistrate within 48 hours of arrest at a so-called *Gerstein* hearing.[18] After arrest, a probable cause determination is *again* made either by a grand jury (if the defendant was not already indicted before arrest) or at a preliminary hearing, leading to a formal charge.

There are also opportunities for both civil and criminal defendants to pursue a dismissal if the relevant pleadings have not make allegations against them that are sufficient to warrant a trial. In the criminal context, a defendant can make a motion for failure to state an offense under Federal Rule of Criminal Procedure 12(b)(3). Similarly, in the civil context, the defendant has an opportunity before trial to move under Federal Rule of Civil Procedure 12(b)(6) for a dismissal based on "failure to state a claim upon which relief be granted," and both parties have opportunities

to move under Rule 12(c) for a Judgment on the Pleadings, or (after discovery) under Rule 56 for Summary Judgment based on the contention that "there is no genuine dispute as to any material fact and the movant is entitled to judgment as a matter of law."

Thus, in both the civil and the criminal contexts, there are a variety of constraints which operate *before* the factfinder applies the relevant standard of proof, and which have an impact on the composition of the population of individuals to which the factfinder applies that standard of proof. Focusing on the criminal context, it is only criminal defendants who have cleared both the probable cause determination (multiple times) and have been unsuccessful at securing a subsequent dispositive ruling who are actually evaluated by the factfinder under the BRD standard. Thus, it is important to evaluate the BRD standard in light of its tendency to produce errors when applied to *that* population of defendants. In other words, the BRD standard is just one of multiple standards that ultimately determine how adjudications are made and thus produce the relevant rates of error; hence, our analysis of the BRD standard must be sensitive to the broader context that produces those adjudications.

There is an analogy here to the way that certain types of medical testing regimes, involving both a "screening" test with a low FNL but a higher FPL, as well as a subsequent "diagnostic" test with a lower FPL.[19] Medical testing regimes also involve a consideration of the tradeoffs between false-positive results and false-negative results, as well as the setting of various thresholds to determine what will count as a positive test result. Thus, in order to evaluate how well a particular testing regime does at balancing those tradeoffs in a manner consistent with values involving public health, it is crucial to consider the *entire* regime. When we focus specifically on the error-related features of a particular diagnostic test, we must remember that the role of that diagnostic test is to deliver results *when applied to a pre-screened population of individuals*—i.e., *after* a screening test has already been used to "filter" the population and thereby isolate the subpopulation to which the diagnostic test will be applied. Analogously, when we focus on the adjudicatory standard that is applied in a particular sort of legal proceeding, we must remember that the role of that standard is to guide adjudications when applied to a population of individuals that have been "pre-screened" by standards such as the probable cause standard, the 12(b)(3) or 12(b)(6) standard, etc.

5 | ERROR AVOIDANCE AND ALLOCATION

Against this background, we are in a position to evaluate various extant proposals about how a standard of proof should be settled upon.

On the Error Comparison View, the goal of choosing a standard of proof should be to achieve the optimal ratio between erroneous findings of no-liability (false negatives) and erroneous findings of liability (false positives); call this ratio the **error-to-error ratio**. Equivalently, the Error Comparison View can be stated as the view that the goal of choosing a standard of proof should be to achieve the optimal proportion of erroneous findings (falses) which are erroneous findings of liability (false positives).

In particular, on the Error Comparison View, if the optimal ratio of erroneous findings of no-liability to erroneous findings of liability is 1:1—i.e, if it is just as undesirable for a responsible person to be found not-liable as it is for a non-responsible person to be found liable—then the standard of proof should be set in a manner that makes these errors equally probable. Similarly, if the optimal ratio of erroneous findings of no-liability to erroneous findings of liability is 1:10—i.e., if it is ten times more undesirable for a non-responsible person to be found liable than it is for a responsible person to be found not-liable, then the standard of proof should be set in a manner

that makes the latter error ten times as likely as the former (i.e., it should be a lot more demanding than in the former case).[20]

One significant challenge for the ERROR COMPARISON VIEW derives from the observation that the "base rate" of responsible individuals in the pool of individuals who go to trial can have a dramatic impact on the value of the error-to-error ratio. For example, if a sufficiently high proportion of defendants who face trial are actually responsible, then even a very low standard of proof will likely lead to a high value of the error-to-error ratio, since there will be a lot of false negatives (since there will be so many responsible individuals to be erroneously found non-liable) and yet there will be very few false positives (since there will be so few non-responsible individuals to be erroneously found liable). Similarly, if a sufficiently high proportion of defendants who face trial are actually non-responsible, then even a very demanding standard of proof will likely lead to a low value of the error-to-error ratio, since there will be very few false negatives (since there will be so few responsible individuals to be erroneously found non-liable) and yet there will be a lot of false positives (since there will be so many non-responsible individuals to be erroneously found liable). Call this the **Base Rate Problem**.[21]

The Base Rate Problem raises serious doubts about the coherence of antecedently deciding on an optimal value for the error-to-error ratio, and then subsequently choosing a standard of proof that produces that value of the error-to-error ratio. Since the error-to-error ratio depends on *both* the standard of proof *and* the base rate of responsible individuals in the relevant population, we apparently have two options. First, we can simply abandon the thesis that producing a particular error-to-error ratio is a central goal in choosing our standard of proof; instead, we should focus on an accuracy-oriented goal that does *not* depend on the base rate. Or second, we can hold onto the ERROR COMPARISON VIEW's commitment to the thesis that the optimal value of the error-to-error ratio is the objective in choosing our standard of proof, and conclude from the Base Rate Problem that the standard of proof should depend on *both* the optimal value of the error-to-error ratio *and* the relevant base rate. On this latter approach, as the base rate changes, the standard of proof should change too so as to maintain the ideal value of the error-to-error ratio: as the base rate of responsible individuals increases, the standard of proof needs to decrease in order to maintain the same value of the error-to-error ratio, and as the base rate decreases, the standard of proof needs to increase.[22] And there is no principled limit on this effect; if the goal is always to achieve the ideal value of the error-to-error ratio, then there is no principled floor on how low the standard of proof could go as the base rate increases.

A significant problem for this second approach is that it is in considerable tension with a strong intuitive resistance to similar uses of statistical information in other aspects of both the law and our intuitive thinking. The Blue Bus and Gatecrasher hypotheticals are two relevant examples here; they each dramatize a resistance to using "naked statistical" base-rate information about a population to establish the liability of an individual member of that population, even in cases where the naked statistical information *appears* to be sufficient to meet a particular probabilistic threshold.[23] There have been a number of attempts to resolve these puzzles by distinguishing "individualized" evidence about a particular allegedly illegal act, which can alone underwrite a finding of liability, from "non-individualized" evidence, which cannot do so, and it is not my goal to evaluate these attempts here.[24] Rather, my point is that we *do* feel a strong resistance to allowing legal liability (as well as reactive attitudes like blame[25]) to be assigned on the basis of statistical facts in closely related contexts, and it is natural to feel the same sort of resistance to a view according to which the standard of proof should vary with the relevant statistical base rate of responsible individuals.

There is a related form of resistance to inferences about individuals on the basis of base rates that manifests elsewhere in the law. One (large and complex) context in which this resistance arises is in discussions of the use of predictive algorithmic tools such as PredPol, SSL, PSA, VPRAI, COMPAS, LSI-R, and SFS in the criminal justice system.[26] Though these tools differ from each other in important ways, they each make use of statistical data involving the conduct of groups of people in order to make predictions about the conduct of individual members of those groups. The use of these tools in the contexts of policing, pretrial release determinations, sentencing, and parole is highly controversial; in particular, the use of these tools raises serious Due Process and Equal Protection concerns, especially as related to minority members of a population.[27] But almost nobody thinks that their use is appropriate during the guilt phase of a trial as evidence of a defendant's criminal conduct, or as evidence of a civil defendant's illegal conduct; even if we stipulate that the algorithm's output is statistically relevant to the defendant's conduct, the intuitive thought is again that the illegal conduct of other members of a defendant's groups should not (at least in general) influence the determination of the defendant's liability.

Moreover, this resistance to reasoning from base rates can extend even to inferences involving *an individual's own* prior conduct. For example, Rule 404 of the Federal Rules of Evidence provides that "evidence of a crime, wrong, or other act is not admissible to prove a person's character in order to show that on a particular occasion the person acted in accordance with the character."[28] A defendant's own prior crimes or wrongs are plausibly *highly* probative of whether the defendant acted in accordance with a bad character on a particular occasion, and there is a natural sense in which that evidence is highly "individualized" to the particular defendant. However, Rule 404 embodies a strong resistance even to the use of statistical evidence about a defendant themself in the service of an inference about their conduct in a particular case. Rather, the Federal Rules insist that the evidential basis for a defendant's liability for a particular act be evidence that bears on that act *in particular*, not evidence that merely establishes that defendant (or anyone else, for that matter) is statistically likely to act illegally *in general*.

Of course, the particular kind of statistical information at issue, and the particular proscribed uses of that information, are different in the contexts of the Blue Bus and Gatecrasher hypotheticals, predictive algorithms, Rule 404, and the Base Rate Problem. But these examples reveal a broad resistance to allowing statistical information to be used in ways that increase the likelihood that a non-responsible individual will be found liable. In light of the Base Rate Problem, this resistance places significant pressure on defenders of the Error Comparison View to explain why that consequence of their view is more palatable here than it is in similar contexts.

Another worry about the Error Comparison View is that, while it is straightforwardly sensitive to the relative *badness* of *incorrect* verdicts, it is equally straightforwardly insensitive to the relative *goodness* of *correct* verdicts, as well as to any comparison of the utilities of good and bad outcomes. The Error Comparison View is plainly grounded in a utilitarian comparison of two possible outcomes of the adjudicatory process: a mistaken finding of liability and a mistaken finding of no-liability. But if we are performing a utilitarian analysis of the possible adjudicatory outcomes, it is not at all obvious why we should ignore the other two possible outcomes: a correct finding of liability and a correct finding of no-liability. If we think that it matters how *bad* it is when an individual is mistakenly found to be liable, then it is natural to also be concerned with how *good* it is when an individual is correctly found to be liable. For instance, correct findings of liability can result in property being restored to its proper owner, dangerous individuals being isolated and rehabilitated, and other beneficial consequences.

This latter worry about the Error Comparison View naturally leads to the Four Utilities View, pioneered by Tribe (1971) and developed by Lillquist (2002) and Laudan & Saunders (2009).

The Four Utilities View proceeds from the thought that the standard of proof ought to be set at the level such that a reasonable factfinder with exactly that level of confidence in the defendant's responsibility would—taking the utilities of all four possible trial outcomes into consideration—be indifferent between a finding of liability and a finding of no-liability; if the factfinder's confidence in the defendant's responsibility exceeds that level then they will opt for a finding of liability, and if it falls short of that level then they will opt for a finding of no-liability. Then, letting

u_{CL} be their utility for a correct finding of liability;

u_{CN} be their utility for a correct finding of no-liability;

u_{ML} be their utility for a mistaken finding of liability;

u_{MN} be their utility for a mistaken finding of no-liability; and

c be their confidence that the defendant is responsible,

the factfinder will be indifferent between a finding of liability and a finding of no-liability precisely when their expected utility of a finding of liability, $c \times u_{CL} + (1 - c) \times u_{ML}$, equals their expected utility of a finding of no-liability, $(1 - c) \times u_{CN} + c \times u_{MN}$. Solving for c, then, we get:

$$c = \frac{u_{CN} - u_{ML}}{u_{CL} + u_{CN} - u_{MN} - u_{ML}} = \frac{1}{1 + \frac{u_{CL} - u_{MN}}{u_{CN} - u_{ML}}}.$$

According to the Four Utilities View, this confidence level should be the threshold at which we set the standard of proof for the proceeding in question. When the factfinder's confidence in the defendant's responsibility exceeds this threshold, their expected utility for a finding of liability is higher than their expected utility for a finding of no-liability, and vice versa when their confidence in the defendant's responsibility falls below this threshold. Thus, using the relevant value of c as a threshold, we can ensure the outcome with the highest expected utility, as judged from the factfinder's perspective.[29]

There are several attractive features of the Four Utilities View. First, the approach fits naturally into a familiar expected-utility-oriented decision-theoretic model;[30] thus, the principles underlying the approach are well-motivated and not idiosyncratic to the particular problem of threshold-selection in legal adjudications. Second, the Four Utilities View straightforwardly addresses the complaint that the Error Comparison View ignores the utilities of correct adjudications, by explicitly including those utilities in the analysis; thus, no unmotivated asymmetry remains between the treatment of correct and incorrect adjudications. Third, the Four Utilities View has at least an apparent advantage over the Error Comparison View when it comes to the Base Rate Problem. After all, the Four Utilities View is sensitive only to the utilities of the four possible trial outcomes and to the factfinder's credence that that defendant is responsible; thus, there is no obvious dependence (as there was for the Error Comparison View) on statistical facts about the distribution of responsible individuals in the population of individuals facing trial.[31]

However, there are also several deep theoretical difficulties for the FOUR UTILITIES VIEW, which arise as well for the ERROR COMPARISON VIEW. First, as I will explore in Section 6, there are a number of important values at stake in the choice of a standard of proof—for example, values related to the incentives and disincentives that the standard imposes on entire populations—that cannot be adequately captured by focusing exclusively on the possible outcomes of adjudications for defendants. Second, there does not appear to be any prospect of faithfully accommodating a notion of *epistemic rights*—i.e., rights to be reasoned about or to not be reasoned about in particular ways—within either the ERROR COMPARISON VIEW or the FOUR UTILITIES VIEW. After all, the central mechanism of both the ERROR COMPARISON VIEW and the FOUR UTILITIES VIEW is an appeal to *utilities*—i.e., an appeal to the (relative or absolute) goodness or badness of various possible *outcomes* of the adjudicatory process. But just as rights in general are notoriously difficult or impossible to capture within a consequentialist framework,[32] it is similarly difficult to see how epistemic rights—if such there be—could be captured within a consequentialist framework like the one underlying both the ERROR COMPARISON VIEW and the FOUR UTILITIES VIEW.[33]

On my view, individuals have a variety of important epistemic rights, some of which are particular to the legal context and others of which are not. One cluster of these rights involves the right not to be reasoned against in certain contexts using particular sorts of inferences—for example, certain kinds of base-rate inferences, or certain inferences involving an individual's "propensity" to act in particular ways,[34] or certain inferences involving individuals' religious beliefs.[35] Another cluster of epistemic rights involves the assumptions that should (or should not) be made about individuals as a starting point for reasoning about them, the ways that various burdens should be allocated when reasoning about them, and the evidential thresholds that should be required for various kinds of conclusions about them.[36]

One function of these rights can be to set a "ceiling" on the value of certain likelihoods, such as the FPL. While there is no requirement that these likelihoods *precisely equal* the "ceiling" value, certain rights require that these likelihoods do not exceed the ceiling value.[37] Decreases below the ceiling might be motivated by a number of different considerations, just as other ways of taking individual or group interests into consideration in general can motivate policy decisions that protect more than just basic rights.[38] Moreover, on my view, defendants are not the only parties that have epistemic rights; in the civil context, the *plaintiff's* right to a proper remedy for legal wrongs committed against them similarly induces a ceiling on the FNL.[39] By contrast, though of course there are weighty *interests* that governments (and and their constituencies) have in seeing to it that responsible individuals are found liable, there are no *individual rights* of the prosecution that are at stake in a criminal proceeding.[40] This asymmetry, I think, is a part of the explanation of why the standard of proof is significantly higher in the criminal context than it is in the civil context.

One important advantage of analyzing accuracy in terms of *likelihoods* such as the FPL and the FNL is that, since these likelihoods are independent of the base rate of responsible individuals in the pool of people who go to trial, the Base Rate Problem does not arise for analyses that focus directly on them. Recall that the FPL is the conditional probability, assuming that an individual is non-responsible, that they will be found liable at trial. This likelihood is unaffected by the proportion of responsible individuals among those who go to trial; that proportion impacts the expected proportion of non-responsible individuals among those who are found liable, but it does not affect the expected proportion of individuals who are found liable among those who are non-responsible.[41]

On the picture I am proposing, the individual epistemic right that gives rise to the ceiling on the FPL is not identical to a putative right not to be falsely found liable; I deny the existence of this latter putative right. Of course an individual's being falsely found liable is a bad consequence, but

on my view *that* bad consequence (or, the reasonable expectation of that bad consequence) can be appropriately weighed, along with other (anticipated) bad consequences, against any number of (anticipated) good consequences in a consequence-oriented analysis. Rather, the epistemic right at issue is an individual's right not to be found liable unless a *sufficient quantity* of evidence can be produced them, where the quantity of evidence that is "sufficient" can depend on the stakes and consequences of the proceeding *for the individual*,[42] but is ultimately to be understood as being an amount of evidence that has a sufficiently low likelihood of being available for presentation against a non-responsible defendant. As long as an individual's epistemic right to that likelihood ceiling is honored, adjustments to the standard of proof that merely weigh their interest in further lowering the FPL along with other interests can be perfectly appropriate.[43]

Compare: my purchase of a lottery ticket confers on me a right to a certain probability of winning the lottery prize. The right at issue here is certainly not a right to *win* the prize, and I have no legitimate complaint arising solely from the fact that I did not win. But I would have a legitimate complaint that would arise if my *probability* of winning were not (at least) as high as advertised—say, if the mechanism underlying the lottery were deliberately manipulated so as to favor the lottery organizer's best friend. For various reasons, the lottery organizers might reasonably and appropriately decide to confer on me some *special* benefit beyond the advertised probability of winning; for example, they might decide to add additional prize money to the prize pool in a manner that increases everyone's chances of winning, or they might decide to send each of the ticket-holders a lottery-themed tote bag so as to assist in the lottery's marketing efforts. But I have no legitimate complaint if the lottery organizers opt for policies that do not include these additional benefits to me, notwithstanding the fact that such policies would further my (legitimate) interests.

Of course, the badness of a particular outcome (such as being mistakenly found to be liable) is not *unrelated* to the individual's epistemic right to a certain likelihood ceiling, any more than the badness of being tortured is unrelated to the right not to be tortured (or than the badness of being deceived or defrauded is unrelated to the right not to unknowingly participate in a rigged lottery). But just as anti-consequentialists reject the thought that the badness of being tortured should be *weighed against* or *compared to* the badness of some other possible outcome(s) in order to explain the prohibition on torture, so too do I find it natural to reject the thought that the badness of an insufficiently demanding standard of proof can be fully accounted for by analyzing the badness of a false positive outcome and weighing that outcome against, or comparing it to, some other possible outcome(s). And yet, this is precisely what both the Error Comparison View and the Four Utilities View do.

6 | POPULATIONS

An important worry for views that focus exclusively on the utilities of trial outcomes, alluded to in the previous section, is that they are insensitive to the various ways that choices about a standard of proof can have impacts on *populations*.

The first such impact is the *deterrence* effect that a standard of proof can have on "primary" (i.e., out-of-court) conduct. Kaplow (2012) provides the most detailed analysis of this phenomenon. Relying on standard approaches to expected-utility theory, Kaplow's starting point is the assumption that an individual's decision to engage in primary behavior which may have legal consequences is influenced by three factors: (1) the probability that the individual's conduct will be the subject of legal action; (2) the probability that, if such legal action is commenced, the individual will be found liable for the conduct in question; and (3) the magnitude of the sanction that

the individual can expect to be subject to if they are found liable.[44] But reducing (increasing) the standard of proof for a particular type of proceeding will surely impact factor (2) by increasing (reducing) the probability that an individual will be found liable, and hence will tend—insofar as they are responsive to this incentive—to decrease (increase) the net expected utility of engaging in the primary conduct in question, which will in turn tend to make them less (more) likely to engage in that conduct. As Kaplow notes, this can impact *both* socially undesirable behavior *and* behavior that is not socially undesirable. For example, a lower standard of proof for a particular crime might both deter criminal behavior—since the would-be criminal has a higher chance of being convicted of the crime—and also "chill" non-criminal (and perhaps even socially desirable) behavior—since the non-criminal conduct in question might be sufficiently similar to criminal behavior that that actor faces increased risks of criminal liability for engaging in it.[45] One noteworthy consequence of this model is that there can be circumstances in which moving to a higher standard of proof can actually *increase*, rather than decrease, the proportion of findings of liability that are mistaken; because of the higher standard of proof, benign primary conduct is incentivized, as a result of which a higher base rate of the cases being tried involve non-responsible defendants, and the proportion of findings of liability that are mistaken can thus increase.[46] Note too that this deterrence phenomenon impacts the *entire population of the jurisdiction*, not just the "pre-screened" population of individuals against whom legal action has been initiated and yet who have failed to secure a pre-trial dismissal.

These deterrent effects—both on legal and illegal primary conduct—cannot be captured by views like the Four Utilities View. For these effects are not consequences of any *particular* result for any *particular* defendant; they are effects on the entire population of people within the relevant jurisdiction, brought about by the setting of the standard of proof *itself*. Of course, results of particular adjudications may *also* have incentivizing or disincentivizing effects; both correct and incorrect individual findings of liability may have (expected) disincentivizing effects on the relevant primary conduct which can be "baked into" u_{CL} and u_{ML}, and both correct and incorrect findings of no-liability may have incentivizing effects which can "baked into" u_{CN} and u_{MN}. But the incentivizing and disincentivizing effects that the adoption of a particular standard of proof *itself* has on a population—independent of any particular application of that standard to any particular defendant, or to any particular trial outcome for any particular defendant—cannot be captured by a model (like the Four Utilities View) that restricts its attention to the values of u_{CL}, u_{CN}, u_{ML}, and u_{MN}.

One noteworthy feature of this disincentivizing effect is that, though it relies on a similar sort of statistical inference to the inferences discussed in Section 5 (involving the Blue Bus and Gatecrasher hypotheticals, predictive algorithms, and Rule 404 of the Federal Rules of Evidence), there is far less intuitive resistance to the use of statistical reasoning in the context of arranging incentives and disincentives for conduct than there is in the context of individual adjudication. In both cases, the statistical inference involves propositions about groups of people—how likely they are to have engaged in particular conduct, how likely they are to modify their conduct in response to various incentives—and in both cases this reasoning about groups leads to important consequences for individual members of the relevant groups. However, whereas I claim that individual epistemic rights are directly threatened by the statistical inferences at stake in the Blue Bus and Gatecrasher hypotheticals, predictive algorithms, and Rule 404, there is no comparably strong intuitive threat to individual rights that is at stake in reasoning about how groups of people are likely to *modify their conduct in the future* in response to a change in the standard of proof.

Second, Parchomovsky & Stein (2010) have identified another way in which individuals' primary conduct can be impacted by features of the legal system. More specifically, Parchomovsky &

Stein argue that evidentiary concerns arising in a variety of fields of law can have a "distortionary" effect on individuals' primary conduct by incentivizing them to act so as to maximize the strength of their legal case, even where the incentivized action is socially undesirable. One example that they give involves law enforcement officers observing a burglary in progress: the officers have an incentive to allow the burglary to progress to the point that a strong case can be built against the suspect in court, even if that means allowing the burglar to break the building's locks and ransack the showcases.[47] A second example involves the law of torts: when a new chemical plant begins to operate and causes damage to nearby property, homeowners have an incentive to "let the harmful effects accumulate" rather than filing a nuisance action too swiftly, so as to collect additional evidence of actual harm and thereby increase the chances of securing an adequate remedy.[48] Though Parchomovsky & Stein do not focus on standards of proof, their central point can be applied to standards of proof as well: if the standard of proof is lower, then both the detectives in the first example and the homeowner in the second example need to collect less evidence in order in order to build an adequately strong case, and hence (other things equal) will tend not to permit the socially undesirable consequences to accrue for as long before taking action to stop them.

A third way in which the standard of proof can impact populations is that, once the primary conduct that potentially gives rise to legal action has already occurred (and hence after the incentives and disincentives that Kaplow focuses on have had their effect), a lower standard of proof can lead to a higher probability that "secondary" legal action—such as a prosecution or a civil lawsuit—will be initiated against a (potential) defendant. When the standard of proof for a type of legal action is lower, it is (other things equal) more likely that the defendant will be found liable, and hence there is a greater incentive to initiate legal action against that person. In some cases this can be a desirable phenomenon: for instance, we want civil plaintiffs who have been wronged by others to have an adequate incentive to initiate suits that may address those wrongs, and an excessively high civil standard of proof can diminish that incentive and thereby discourage meritorious suits. On the other hand, a low civil standard of proof can encourage frivolous "strike suits" aimed merely at securing a settlement from a defendant who would prefer to avoid the hassle and expense of defending themself against the suit, even though they would almost certainly prevail at trial. Relatedly, higher standards of proof (especially applied to awards of punitive damages, discussed in Section 7 below) can help to keep insurance costs low for entities that insure themselves against civil damages, which can also reduce the litigation costs that such entities pass along to consumers. Similarly, in the criminal context, the high standard of proof discourages prosecutions against defendants unless the prosecution reasonably expects to be able to build an extremely strong case against the defendant; the result, for both better and worse, is that many defendants against whom a merely moderately strong case can be built are never charged with crimes.

Similar remarks apply to the initiation of *investigations* that might precede formal legal action. For instance, it may not be worth launching an investigation into a particular matter if the standard of proof is so high that the anticipated fruits of the contemplated investigation cannot be reasonably expected to clear that standard; by contrast, a lower standard incentivizes more investigations, since it is more likely that they will result in successful legal actions. Incentivizing more investigations can have positive and negative effects: more investigations lead to more discovery (and more disincentivization) of wrongdoing, but they also result in expenditures of time and other resources, as well as more non-responsible individuals bearing the costs of having their affairs investigated.

Once legal action is initiated against the defendant, the standard of proof that would govern at trial looms over settlement and plea negotiations, impacting all parties' estimates of their likely success at trial and thus of the expected outcome of a trial. The result is that higher standards

of proof, other things equal, tend to lead to more defendant-favorable settlements and pleas; if a defendant merely has to raise a reasonable doubt about their responsibility at trial to avoid liability, their negotiating position is much stronger than it is when a preponderance is sufficient for a finding of liability. Relatedly, if the defendant is sufficiently confident that the strength of the case against them is inadequate to clear the standard of proof, then they may choose to go to trial rather than to accept *any* plea or settlement offer that is likely to be made.

Once trial starts, the standard of proof more explicitly governs the outcome of the proceeding, which can under certain circumstances be determined by the judge (acting as trier of law) rather than by the factfinder. The reason is that there are a variety of opportunities for parties to take the trial out of the hands of the factfinder, by moving for a dispositive ruling by the judge either before or after the factfinder has rendered a verdict. Unlike the logically independent standards discussed in Section 4 (the probable cause standard, the standard under Federal Rule of Criminal Procedure 12(b)(3), and the standard under Federal Rule of Civil Procedure 12(b)(6)), the standards for these determinations are parasitic on the standard of proof that the jury will apply at the end of the trial. For example, in the civil context, there are several opportunities for parties to move for Judgment as a Matter of Law under Rule 50 of the Federal Rules of Civil Procedure, which will be granted "[i]f a party has been fully heard on an issue and the court finds that a reasonable jury would not have a legally sufficient evidentiary basis to find for the party on that issue"; clearly, "legally sufficient evidentiary basis" here is shorthand for "evidentiary basis that is legally sufficient to meet the relevant standard of proof." Similarly, after a criminal trial has begun, there are opportunities for the defendant to raise a motion for Judgment of Acquittal under Federal Rule of Criminal Procedure 29, which will be granted if "the evidence is insufficient to sustain a conviction"; again, "insufficient to sustain a conviction" here means "insufficient to sustain a conviction by the relevant standard of proof." As a result, the choice of a standard of proof to be applied by the factfinder at the end of a trial induces other decisions about how long the trial will continue under various conditions, whether the factfinder will be permitted to render a verdict, and whether the factfinder's verdict will be set aside by the presiding judge, all of which can impact the interests of judges, juries, litigants in other cases on the court's docket, and the public. Relatedly, the choice of standard also impacts the way that a trial verdict will be reviewed by appellate courts; for example, the question of whether an error that was committed at trial was "harmless" or not often makes explicit appeal to the standard of proof at trial.[49]

A fourth and final population-level value that is impacted by the choice of a standard of proof involves the game-theoretic balance between the opposing sides in litigation. Posner (1999), for instance, has argued that there is in general an important game-theoretic difference between criminal and civil litigation: "The government has enormous prosecutorial resources [that it can] allocate…across cases as it pleases, extracting guilty pleas by threatening to concentrate its resources against any defendant who refuses to plead and using the resources thus conserved to wallop the occasional defendant who does invoke his right to a trial."[50] Posner goes on to suggest that the extremely demanding BRD standard in criminal cases can be understood as a "partial offset (like the provision of counsel to indigent defendants) to the inequality of the parties' resources for gathering and presenting evidence."[51] Of course, one consequence of this "offset" is that criminal defendants who *can* devote significant resources to their defense enjoy disproportionate advantages: they get the benefit of the same offset as every other criminal defendant in the form of the BRD standard, but in their case the offset is less necessary because of the reduced asymmetry of power between the prosecution and the defense. By contrast, though of course it is often true that there are asymmetries of power and resources between opposing parties in civil litigation, these asymmetries are less systematic and stable, and can in some cases favor the defendant rather than

the plaintiff. Moreover, even where a civil plaintiff does enjoy a large power or resource advantage over a particular defendant, it is not typically the case that well-resourced plaintiffs have the same incentives as a prosecutor's office to obtain settlements from large numbers of similarly-situated defendants. Thus, there is no comparable need for the standard of proof to be used as a partial offset in order to maintain the appropriate game-theoretic balance in the context of civil litigation.

Again, since u_{CL}, u_{CN}, u_{ML}, and u_{MN} reflect only the utilities of various *trial outcomes for a particular defendant*, there is no reasonable prospect for the FOUR UTILITIES VIEW to fully account for any of the phenomena discussed in this section. Each of these phenomena involves effects on individuals who never actually face trial, or conduct that occurs outside of a trial, or effects that impact trials but cannot be accounted for solely with reference to the values of the four trial outcomes.

7 | TIGHTER AND LOOSER CONSTRAINTS

On the approach I have been developing, the standard of proof is constrained first by epistemic rights such as the defendant's right to a FPL that is adequately low in light of the consequences of liability, and a civil plaintiff's right to a FNL that is adequately low in light of their interests. In many cases, these epistemic rights may constrain the standard of proof so tightly that other effects—such as the population-level effects discussed in Section 6– - do not have "room" to make a difference to the standard of proof. The BRD standard in criminal contexts is a good example here; the defendant's epistemic right to a low FPL is at its strongest when involuntary confinement and other criminal sanctions are at stake, which leaves very little room for other non-rights-based considerations to motivate an upward shift in the standard of proof. By contrast, in other situations, the epistemic rights at stake more significantly underdetermine the standard of proof, in which case considerations like the ones addressed in Section 6 have more room to operate.

A useful illustration involves the "clear and convincing evidence" standard that serves as an intermediate standard of proof, sitting in between the preponderance standard and the BRD standard.[52] The U.S. Supreme Court has held that proof by this intermediate standard is required, either by the U.S. Constitution or by relevant federal statutes, in cases involving deprivations of individual rights outside of the criminal context, including civil commitment,[53] decisions to terminate life,[54] termination of parental rights,[55] denaturalization,[56] and deportation.[57]

The intermediate standard has also been commonly applied in "civil cases involving allegations of fraud or other quasi-criminal wrongdoing by the defendant," on the grounds that "[t]he interests at stake in those cases are deemed to be more substantial that mere loss of money" because they risk erroneously tarnishing the defendant's reputation more profoundly than would an ordinary civil judgment.[58] In these kinds of cases, it is natural to understand the intermediate standard, and hence the intermediate ceiling on the FPL, as being closely connected with the intermediate significance of the individual rights at stake in these types of proceedings: less significant than those at stake in criminal sanctions, but more significant than "mere" financial interests.

However, there is considerable variation among U.S. jurisdictions with regard to application of the intermediate standard to other kinds of determinations, where it is often far less clear that the motivation is to protect important individual rights. For example, states vary significantly in the standard that is applied to punitive damage awards, with many states applying the intermediate standard to punitive damages, even where the preponderance standard applies to the underlying civil claim.[59] And although *one* rationale for punitive damages awards—non-criminal punishment of the defendant for wrongdoing—implicates some of the individual rights discussed above,

others do not. For example, deterrence is central to many awards of punitive damages: the defendant knew about some potential problem but took inadequate steps to address it because, they reasoned, it is cheaper to pay out occasional civil damage awards as a "cost of doing business," rather than investing in a genuine solution to the problem. Insofar as courts and legislatures want to disincentivize this sort of reasoning (both by the defendant and by similarly situated others), they will be open to allowing punitive damages so as to increase the expected costs of inadequately addressing problems that may give rise to legal liability. And, the lower the standard of proof that is associated with such punitive damages awards, the more effective punitive damages will be at deterring conduct that might give rise to punitive damages. (And note again that, as discussed in Section 6, deterrence-oriented objectives raise fewer concerns about individual rights than other objectives do.) Thus, a natural reconstruction of at least some of the variation in standards applied to punitive damages is that different jurisdictions are coming to different reasonable conclusions about whether and by how much to lower the FPL and FNL below the rights-preserving ceiling.

The intermediate standard has also been applied in a variety of other situations where there is thought to be some special danger of deception, or where a particular type of claim or defense is disfavored on policy grounds. For example, suits for specific performance of an oral contract have commonly been subjected to the intermediate standard, on the grounds that oral contracts are particularly subject to fraud and misunderstanding.[60] Proceedings to set aside or modify written transactions or official acts on grounds of fraud or mistake have also been subjected to the intermediate standard; at least part of the rationale here is the goal of encouraging trust in these transactions, which might be undermined if a preponderance were all that was required to set them aside.[61] The intermediate standard has also been applied to a wide variety of issues in the law of trusts and estates, including: (a) proof of testamentary intent in applications of the "harmless error" rule where the Wills Act formalities were not precisely followed; (b) proof that a testator's intent to revoke a will was conditional on the validity of a new will, in applications of the doctrine of dependent relative revocation; (c) proof of the contents of a lost but unrevoked will; (d) proof of the existence of a contract not to revoke a will; and (e) proof of a testator's mistake, in suits to reform a will for mistake.[62] Surely, these applications of the intermediate standard each raise a unique set of issues, but one feature that is common to each of them is the danger of deception and fraud that arises whenever the key witness (i.e., a deceased testator) is unavailable, and interested parties are the primary sources of relevant evidence. For instance, if a mere preponderance were sufficient to prove the contents of a lost but unrevoked will, then a disinherited family member who found a will after the testator's death would often have strong incentive to destroy the will and lie about its contents. Of course, there are also certain property rights at stake here, both for the testator and for a legitimate devisee under a lost but unrevoked will, and perhaps too high a standard (say, the BRD standard) would risk violating those rights. But many jurisdictions have come to the conclusion that imposing the intermediate standard on certain types of claims is both consistent with the individual rights at stake and well-justified on policy grounds. And, once more, it is worthy of note that many jurisdictions are comfortable using the intermediate standard as a means of discouraging disfavored primary and secondary conduct, quite apart from the goal of allocating the risk of error in the adjudication of claims at trial.

ENDNOTES

[1] I do not mean to suggest that the notion of accuracy characterized here is the only legitimate notion of accuracy that matters in the law, or that nothing other than accuracy impacts the normative status of an adjudication. For example, if the evidence is insufficient to establish the defendant's legal liability in either a criminal or civil context, there is a perfectly good sense in which it is not an "error" for them to be found not to be liable at trial,

even if they *actually* engaged in the proscribed activity; regardless of what the defendant did or didn't actually do, there is a perfectly good sense in which the correct verdict is the one supported by the evidence, even if that verdict doesn't match the facts. However, there is *also* a perfectly good sense in which the imagined result *is* an error, since by hypothesis the defendant really did engage in the proscribed activity, and it is this sense of "error" that I am focusing on here. Indeed, the standard that we set for the evidence to be "sufficient" is precisely the focus of this paper; one important question here is how to set the evidential standard so that a body of verdicts which are correct in the evidential sense tends to contain the lowest proportion of verdicts which are errors in the factual sense.

[2] *United States v. Havens*, 446 U.S. 620, 626 (1980).

[3] *See, e.g.*, *Arizona v. Fulminante*, 499 U.S. 279, 308 (1991).

[4] *See, e.g.*, Kaplow (2012) at 757. Johnson King (forthcoming) argues that we do not have good justification to accept this assumption, since we have no epistemic access to who is *actually* responsible or non-responsible, apart from our epistemic access to who *seems* responsible or non-responsible to factfinders. This raises an interesting skeptical challenge which I cannot adequately address here.

[5] There are some reasons to be hesitant about this "one-dimensional" approach to standards of proof, including reasons that arise in the so-called "Blue Bus" and "Gatecrasher" hypotheticals, discussed below.

[6] Burdens come in two different forms in the law: the "burden of persuasion," which identifies the party who must ultimately establish some fact (by the relevant standard of proof), and the "burden of production," which identifies the party who must introduce evidence of a certain sort in order to avoid an adverse directed verdict. Again, the differences here are subtle, as are the connections between both sorts of burdens and the standard of proof.

[7] Blackstone (1893) at 358.

[8] "Again, every one of us would rather acquit a guilty man as innocent than condemn an innocent man as guilty, in a case where a man was accused of enslaving or murder. For in each of these cases if the charges were true we should prefer to vote for their acquittal on the charges against them, rather than to vote for their condemnation, if the charges were untrue. For when there is any doubt one should choose the lesser of two evils. For it is a serious matter to decide in the case of a slave that he is free; but it is much more serious to condemn a free man as a slave." Aristotle (1937) at 144-45.

[9] "… for it is better five guilty persons should escape unpunished, than one innocent person should die." Hale (1778) at 289.

[10] "Indeed I would rather wish twenty evildoers to escape death through pity, than one man to be unjustly condemned." Fortescue (1967) at 63.

[11] "[I]t is better and more satisfactory to acquit a thousand guilty persons than to put a single innocent man to death once in a way." Maimonides (1967) at 270.

[12] 397 U.S. 358, 371-72 (1970).

[13] Of course, there may be ways of resisting the argument that, in practice, the requirement of sworn testimony really does reduce both the FPL and the FNL. But the more general point still stands, and could as easily be made with an example like Federal Rule of Evidence 402's requirement that admissible evidence be relevant to a fact of consequence, or Federal Rule of Evidence 403's directive that the court weigh, *inter alia*, the danger of unfair prejudice resulting from the introduction of a piece of evidence against its probative value.

[14] DeKay (1996) and Laudan (2006) similarly observe that, other things equal, higher standards of proof tend to lead to more erroneous acquittals and fewer erroneous convictions.

[15] *See generally* Kaplow (2013).

[16] See, e.g., *People v. Ayala*, 770 P.2d 1265 (Colo. 1989). It is often made explicit that, in applying this standard, "the evidence must be viewed in the light most favorable to the prosecution, and all inferences must be resolved in favor of the prosecution," which has the effect of significantly relaxing the standard. Also, note that there is no general requirement that evidence used to support the probable cause determination be admissible; *see, e.g.*, Federal Rule of Evidence 1101(d).

[17] *See generally* Cohen et al. (2019) chapter 6 §C.7; *see also* Meyn 2014.

[18] *See Gerstein v. Pugh*, 420 U.S. 103 (1975).

[19] Epidemiologists generally speak in terms of "sensitivity" and "specificity"; a sensitive test is very likely to "notice" a genuine instance of a condition when the test is confronted with one, whereas a specific test is unlikely

to be positive unless the patient has the condition being tested for. Screening tests are typically sensitive but not very specific, whereas diagnostic tests are typically more specific. *See generally* Rothman (2012) chapter 13.

[20] Of course, making these judgments with precision is unrealistic, and standards like the BRD standard are notoriously (and likely deliberately) resistant to probabilistic precisification. But the more general framework here—according to which a lower value of the error-to-error ratio motivates a more demanding standard of proof—does not essentially depend on the feasibility of calculating the error-to-error ratio with mathematical precision.

[21] Versions of this worry have been presented in Allen (1977), Bell (1987), Connolly (1987), DeKay (1996), Johnson King (forthcoming), Lillquist (2002), and Laudan (2006), among others. DeKay's statement of the worry is that "[the] standard of proof employed by the jury does not, by itself, determine the ratio of judicial errors . . . [which] also depends on the prior odds of guilt and the accuracy of the jury" (DeKay (1996) at 126).

[22] *See* Ribeiro (2019) for a discussion and defense of varying standards of proof.

[23] In the Blue Bus hypothetical, the puzzle is why we are unwilling to assign civil liability to the Blue Bus Company based solely on evidence that the plaintiff was hit by a bus in a town in which Blue Bus Company operates 80% of the buses; intuitively, it seems as though this statistical information establishes an 80% probability that Blue Bus Company was responsible (which seems to clearly meet the preponderance standard), and yet we are unwilling to assign civil liability to the Blue Bus Company on that basis alone. Similarly, in the Gatecrasher hypothetical, we are unwilling to assign criminal liability to a particular defendant based solely on the statistical fact that they were found in an area containing 100 people, 99 of whom had not paid for entrance to an event; again, though it seems as though this does establish a probability of 99% that this defendant committed the crime of gatecrashing, and that this clearly meets the BRD standard, still we are hesitant to assign criminal liability on the basis of this statistical information alone.

[24] For discussions of statistical evidence and of the Blue Bus and Gatecrasher hypotheticals, *see, e.g.*, Blome-Tillmann (2015), Bolinger (2020), Buchak (2014), Enoch et al. (2012), Gardiner (2019), Jackson (2020), Littlejohn (2020), Redmayne (2008), Staffel (2016), Smith (2018), and Thomson (1986).

[25] Buchak (2014).

[26] *See generally* O'Neil (2017), Yang & Dobbie (2020).

[27] See Yang & Dobbie (2020).

[28] Federal Rules of Evidence 404(b).

[29] *See* Clermont (2013) at 30; *see also* Laudan & Saunders (2009) at 3–4.

[30] *See generally* Steele & Stefánsson (2020).

[31] Of course, the question does then naturally arise of whether the *value of c*—i.e., the value of the factfinder's credence that the defendant is responsible—is itself sensitive to the base rate of responsible individuals in the relevant population; if it is, then the Base Rate Problem re-arises (albeit in a slightly different form) for the Four Utilities View. But the Four Utilities View has more resources here than the Error Comparison View does, since it is possible to set and apply thresholds that are to be applied by factfinders against the background of a broader body of rules of evidence, including ones that forbid certain types of base rate information from explicitly figuring into the factfinder's credence that the defendant is responsible. (For example, *see* discussion above of the Blue Bus and Gatecrasher hypotheticals, predictive algorithms, and Rule 404.) Of course, there are some types of base rate evidence that it is entirely proper for a factfinder to take into consideration; for example, a defense that appeals to a wildly improbable coincidence such as two randomly selected individuals sharing a DNA profile might reasonably be rejected by a factfinder on the basis of the low base rate of such a coincidence. Moreover, the suggestion here is not that, in light of particular doctrines in the law of evidence, no factfinder will ever improperly appeal to base rate information; implicit bias and other forms improper statistical reasoning are almost persist even when factfinders are properly instructed on the law of evidence. Rather, the suggestion is that whereas the counterintuitive sensitivity of thresholds to base rate is unavoidably "baked into" the Error Comparison View, the Four Utilities View has resources to limit that sensitivity. In that context, it is possible on the Four Utilities View for the question of where to set the standard of proof to be independent of the base rate of responsible individuals; by contrast, the logical and mathematical structure of the Error Comparison View makes that impossible.

[32] Some consequentialists—e.g., Singer (1974), Unger (1996), and Norcross (1997)—advocate "bullet-biting" solutions, whereas others—e.g., Sen (1982), Broome (1991), and Portmore (2001, 2003)—have proposed "agent-relative" versions of consequentialism that are designed to accommodate thought and talk about rights. Indirect consequentialists—e.g., Singer (1961), Brandt (1992), Gert (2005), and Rawls (1955)—also have strategies

for accommodating thought and talk about rights into a broadly consequentialist framework. See also Muñoz (2021).

[33] Ronald Dworkin (1981) has emphasized the "moral harm" associated with being falsely convicted which "will escape the net of any utilitarian calculation." In addition, several other theorists have raised related worries about using utilitarian calculations to weigh the benefits of accurate verdicts against the costs of inaccurate ones. *See, e.g.*, Stein (2005), Tribe (1971), and Walen (2015).

[34] Rule 404(b) of the Federal Rules of Evidence formalizes the prohibition on "propensity inferences" involving an individual's character in the legal context.

[35] Rule 610 of the Federal Rules of Evidence formalizes the prohibition on appealing to evidence of a person's religious beliefs in order to attack (or support) their credibility.

[36] The U.S. Supreme Court has explicitly recognized the connection between individual rights and the thresholds imposed by standards of proof: "In cases involving individual rights, whether criminal or civil, [t]he standard of proof [at a minimum] reflects the value society places on individual liberty." *Addington v. Texas*, 441 U.S. 418, 425 (1979).

[37] Of course, a "ceiling" on the false positive likelihood could easily be re-conceptualized as a "floor" on the true positive likelihood.

[38] Similarly, Ronald Dworkin has argued that rights are to be given lexical priority over any non-rights-based consideration, but that non-rights-based considerations can be morally significant in cases where the relevant rights fail to determine a unique outcome. *See, e.g.*, Dworkin (1973, 1975, 1981), and (1986).

[39] Both the 7th Amendment to the U.S. Constitution and Rule 38 of the Federal Rules of Civil Procedure protect the right to a jury trial in certain civil matters, which encompasses a right to trial on those matters. I am here suggesting that the plaintiff's right to a civil trial, in turn, encompasses a right to a FNL that is no higher than the relevant ceiling.

[40] One natural thought here is that perhaps the *victim* of a crime (at least for crimes with clearly identifiable victims) also have individual rights which induce a ceiling on the FNL in criminal cases. However, I do not think this is correct. Criminal victims certainly do have certain carefully circumscribed rights in criminal cases, including the right under Federal Rule of Evidence 412 not to have evidence introduced against them (under certain circumstances) in order to prove their "sexual predisposition." But I am not persuaded that victim's rights have a distinctive role to play *in setting the standard of proof*. The tradition here, with which I am largely sympathetic, is to see criminal proceedings as actions in which the only parties are the government and the defendant, and in which the victim's preferences are ultimately not dispositive, although they are often (appropriately) taken very seriously. And, unlike civil plaintiffs or criminal prosecutors, victims do not in general have the right to decide whether charges are filed, to approve or veto a plea deal, to testify, or to direct litigation goals or strategy; I think it would be somewhat anomalous to think that victims have epistemic rights that further constrain the standard of proof, even after the defendant's rights have imposed separate (and very stringent, on my account) constraints.

[41] This is a familiar point from the literature on Bayesian inference, and is closely related to idea that, while priors reflect subjective initial doxastic attitudes toward theories, likelihoods capture an objective feature of the relationship between theory and evidence.

[42] So, for example, stakes and consequences for other people, or for society at large, are not yet getting into the analysis, as they would for a straightforward consequential analysis.

[43] I am broadly sympathetic to the view about the criminal standard of proof that is endorsed in Laudan (2006), though that view is presented in a somewhat misleading manner. Laudan focuses on the "fate of the innocent" and endorses the view that the standard of proof should be configured so as to produce the socially ideal ratio of correct findings of no-liability to erroneous findings of liability, which he calls m; the value of m is settled by the the choice of FPL. Laudan (2006) further holds that this approach should be supplemented with a "side-constraint" generated by the value of the error-to-error ratio: "the socially settled value of [the error-to-error ratio] acts as a side constraint on further jiggering with m." (p. 75.) In particular, Laudan claims, "we should insist that the system commit as many [false findings of no-liability] as are necessary in order to preserve m, but no more than that." (p. 75) However, talk about the "ideal" value of m is misleading; m is the ratio of a good thing to a bad thing, so the ideal value is infinite. (By contrast, talk of the "ideal" value of the error-to-error ratio at least makes sense, since it is the ratio of two bad things.) And once we choose a particular value of m, it is not obvious that there are any *further* steps that we can take to impact the value of the error-to-error ratio. For this reason, I prefer to talk about a "ceiling" on the value of the FPL; as long as it goes no higher than that

ceiling value, then "jiggering" of lots of different sorts—and not just to the value of the error-to-error ratio—is perfectly appropriate.

[44] Kaplow (2012) at 754.

[45] Kaplow (2012) at 738–39.

[46] Kaplan (2012) at 789–91. Of course, as a result of the higher standard of proof, illegal primary conduct is also incentivized, which puts downward pressure on the base rate of benign acts among the cases being tried. But, as Kaplan points out, "if the heightened evidence threshold reduces the chilling of benign acts relatively more than it reduces the deterrence of harmful acts," then the net effect on the base rate of benign acts among the cases being tried will be positive.

[47] Parchomovsky & Stein (2010) at 520–21.

[48] Parchomovsky & Stein (2010) at 520.

[49] *See, e.g., Chapman v. California*, 386 U.S. 18 (1967).

[50] Posner (1999) at 1505.

[51] Posner (1999) at 1505.

[52] Formulations of this intermediate standard vary significantly. The Supreme Court in *Addington v. Texas* characterized it this way: "The intermediate standard, which usually employs some combination of the words 'clear,' 'cogent,' 'unequivocal,' and 'convincing,' is less commonly used, but nonetheless 'is no stranger to the civil law.' " 441 U.S. 418, 424 (1979).

[53] *See, e.g.*, *Addington*, 441 U.S. at 427, 431.

[54] *See, e.g.*, *Cruzan v. Mo. Dep't of Health*, 497 U.S. 261, 265 (1990)

[55] *See, e.g., Santosky v. Kramer*, 455 U.S. 745, 750, 769 (1982).

[56] *See, e.g.*, *Chaunt v. United States*, 364 U.S. 350, 353 (1960); *Schneiderman v. United States*, 320 U.S. 118, 125, 159 (1943).

[57] *See, e.g.*, *Woodby v. INS*, 385 U.S. 276, 277, 285 (1966)

[58] *Addington*, 441 U.S. at 424.

[59] *See* "Punitive Damages," 0020 Surveys 25, 50 State Statutory Surveys: Civil Laws: Remedies.

[60] *See* Kaye et al. (2014) at 723–24.

[61] Kaye et al. (2014) at 723–24.

[62] *See, e.g.*, UPC §2-503 (harmless error), *Dan v. Dan*, 288 P.3d 480 (Alaska 2012) (lost will); Restatement (Third) of Property, Wills and Other Donative Transfers §4.3 (dependent relative revocation); *Keith v. Lulofs*, 724 S.E.2d 695 (Va. 2012) (contract not to revoke); UPC §2-805 (reformation for mistake).

REFERENCES

Allen, A. (1977). The restoration of in re winship. 76 Mich. L. Rev. 30.

Aristotle. (1937). *Problems* (W.S. Hett trans.). Cambridge: Harvard University Press.

Bell, R. (1987). Decision theory and due process: A critique of the supreme court's lawmaking for burdens of proof. *Journal of Criminal Law and Criminology*, *78*, 557–585.

Blackstone. (1893). *Commentaries on the laws of England*. Philadelphia: J.B. Lippincott Co.

Blome-Tillmann, M. (2015). Sensitivity, causality, and statistical evidence in courts of law. *Thought*, *4*(2), 102–12.

Brandt. (1992). *Morality, utilitarianism, and rights*. Cambridge: Cambridge University Press.

Bolinger, R. (2020). Varieties of moral encroachment. *Philosophical Perspectives*, *34*(1), 5–26.

Broome, J. (1991). *Weighing Goods*. Oxford: Basil Blackwell.

Buchak, L. (2014). Belief, credence, and norms. *Philosophical Studies*, *169*(2), 285–311.

Clermont, K. (2013). *Standards of Decision in Law*. Durham: Carolina Academic Press.

Cohen, N., Adelman, S., Abramson, L., O'Hear, M., & Logan, W. (2019). *Criminal Procedure: The Post-Investigative Process, 5th Ed*. Durham: Carolina Academic Press.

Connolly, T. (1987). Decision theory, reasonable doubt, and the utility of erroneous acquittals. *Law and Human Behavior*, *11*(2), 101–112.

DeKay, M. (1996). The difference between blackstone-like error ratios and probabilistic standards of proof. *Law & Social Inquiry*, *21*, 95–132.

Dworkin, R. (1973). Taking Rights Seriously. In A.W.B. Simpson (Ed.), *Oxford Essays in Jurisprudence, Second Series*, 202. Oxford: Clarendon Press.

Dworkin, R. (1975). Hard Cases. Harvard Law Review 88, 1057–1110.
Dworkin, R. (1981a). Principle, Policy, Procedure. In C. Tapper (Ed.), *Crime, Proof, and Punishment*. 193.
Dworkin, R. (1981b). Is there a right to pornography? "*Oxford Journal of Legal Studies*, *1*, 177–213.
Dworkin, R. (1986). *Law's Empire*. London: Fontana.
Enoch, D., Spectre, L., & Fisher, T. (2012). Statistical evidence, sensitivity, and the legal value of knowledge. *Philosophy and Public Affairs*, *40*(3), 197–224.
Fortescue, J. (1567). *A Learned Commendation of the Politique Laws of England* (Robert Mulcaster trans.) 63.
Gardiner, G. (2019). The reasonable and the relevant: Legal standards of proof. *Philosophy and Public Affairs*, *47*(3), 288–318.
Gert, B. (2005). *Morality: Its Nature and Justification*. New York: Oxford University Press.
Hale, M. (1778). *The History of the Pleas of the Crown* (George Wilson ed.). London: T. Payne.
Jackson, E. (2020). Belief, credence, and evidence. *Synthese*, *197*(11), 5073–5092.
Johnson King, Z. (forthcoming). The Trouble with Standards of Proof. Forthcoming in *Synthese*.
Kaplow, L. (2012). Burden of Proof, 121 Yale L.J. 738.
Kaplow, L. (2013). Multistage Adjudication. 126 Harv. L. Rev. 1179.
Kaye, D., Broun, K., Dix, G., Swift, E., Roberts, E., Imwinkelried, E., & Mosteller, R. (2014). *McCormick's Evidence, 7th* (Hornbook Series).
Laudan, L. (2006). *Truth, Error, and Criminal Law*. New York: Cambridge University Press.
Laudan, L., & Saunders, H. (2009). Re-Thinking the Criminal Standard of Proof: Seeking Consensus About the Utilities of Trial Outcomes (March 29, 2009). Available at SSRN: https://ssrn.com/abstract=1369996 or http://doi.org/10.2139/ssrn.1369996
Lillquist, E. (2002). Recasting Reasonable Doubt: Decision Theory and the Virtues of Variability, 36 U. C. Davis L. Rev. 85.
Littlejohn, C. (2020). Truth, knowledge, and the standard of proof in criminal law. *Synthese*, *197*, 5253–5286.
Maimonides, M. (1967). *The Commandments* (C. Chabel trans.) 270.
Meyn, I. (2014). The Unbearable Lightness of Criminal Procedure, 42 Am. J. Crim. L. 39.
Muñoz, D. (2021). The rejection of consequentializing. *Journal of Philosophy*, *118*(2), 79–96.
Norcross, A. (1997). Comparing harms: Headaches and human lives. *Philosophy and Public Affairs*, *26*, 135–67.
O'Neil, C. (2017). *Weapons of Math Destruction*. New York: Broadway Books.
Parchomovsky, G., & Stein, A. (2010). The Distortionary Effect of Evidence on Primary Behavior. 124 Harv. L. Rev. 518.
Portmore, D. (2001). Can an act-consequentialist theory be agent-relative? *American Philosophical Quarterly*, *38*, 363–77.
Portmore, D. (2003) Position-relative consequentialism, agent-centered options, and supererogation. *Ethics*, *113*, 303–32.
Posner, R. (1999). An Economic Approach to the Law of Evidence. 51 Stan. L. Rev. 1477.
Rawls, J. (1955). Two concepts of rules. *Philosophical Review*, *64*, 3–32.
Redmayne, M. (2008). Exploring the proof paradoxes. *Legal Theory*, *14*(4), 281–309.
Ribeiro, G. (2019). The Case for Varying Standards of Proof. 56 San Diego L. Rev. 161.
Rothman, K. (2012). *Epidemiology: An introduction, 2nd edition*. New York: Oxford University Press.
Sen, A. (1982). Rights and agency. *Philosophy and Public Affairs*, *11*(1), 3–39.
Singer, P. (1974). Sidgwick and reflective equilibrium. *Monist*, *58*, 490–517.
Singer, M. (1961). *Generalization in Ethics*. New York: Knopf.
Smith, M. (2018). When does evidence suffice for conviction? *Mind*, *127*(508), 1193–1218.
Staffel, J. (2016). Beliefs, buses and lotteries: Why rational belief can't be stably high credence. *Philosophical Studies*, *173*(7), 1721–1734.
Steele, K., & Stefánsson, H. (2020). Decision Theory. In the Stanford Encyclopedia of Philosophy (Winter 2020 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2020/entries/decision-theory/.
Stein, A. (2005). *Foundations of Evidence Law*. Oxford: Oxford University Press.
Thomson, J. (1986). Liability and individualized evidence. *Law and Contemporary Problems*, *49*(3), 199–219.
Tribe, L. (1971). Trial by Mathematics: Precision and Ritual in the Legal Process. 84 Harv. L. Rev. 1329.
Unger, P. (1996). *Living High and Letting Die*. New York: Oxford University Press.
Walen, A. (2015). Proof Beyond Reasonable Doubt: A Balanced Retributive Account. 76 La. L. Rev. 355.

Yang, C., & Dobbie, W. (2020). Equal Protection Under Algorithms: A New Statistical and Legal Framework. 119 Mich. L. Rev. 291.

How to cite this article: Kotzen, M. (2021). Standards and values. *Philosophical Issues*, 1–21. https://doi.org/10.1111/phis.12208