Adjusting Inferential Thresholds to Reflect Nonepistemic Values

Kim Kaivanto and Daniel Steel*†

Many philosophers have challenged the ideal of value-free science on the grounds that social or moral values are relevant to inferential thresholds. But given this view, how precisely and to what extent should scientists adjust their inferential thresholds in light of nonepistemic values? We suggest that signal detection theory provides a useful framework for addressing this question. Moreover, this approach opens up further avenues for philosophical inquiry and has important implications for philosophical debates concerning inductive risk. For example, the signal detection theory framework entails that considerations of inductive risk and inferential-threshold placement cannot be conducted in isolation from base-rate information.

1. Introduction. Inferential thresholds specify what evidence should be taken as sufficient for accepting, asserting, or acting on a claim in a context. An inferential threshold might be a level of statistical significance, such as 0.05, or that a certain concentration of antibodies in a blood sample should be taken to indicate that a person is HIV positive. Literature on inductive risk has discussed the role of values of a social or ethical character in decisions about where inferential thresholds should be placed, with particular attention being given to the implications of this issue for the ideal of value-free science (see Churchman 1948; Rudner 1953; Jeffrey 1956; Levi 1962; Hempel 1965; Douglas 2000, 2009; Wilholt 2009; Steel 2010; Betz 2013). However, this philosophical debate has been mostly carried out in an informal manner. Those who argue that nonepistemic values are relevant to scientific

Received May 2017; revised May 2018.

*To contact the authors, please write to: Kim Kaivanto, Lancaster University, Lancaster LA1 4YX, United Kingdom; e-mail: k.kaivanto@lancaster.ac.uk. Daniel Steel, University of British Columbia, Vancouver, BC V6T 1Z3, Canada.

†The authors would like to thank the editor, two anonymous reviewers, and seminar participants at the Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München. The usual disclaimer applies.

Philosophy of Science, 86 (April 2019) pp. 255–285. 0031-8248/2019/8602-0004$10.00
Copyright 2019 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

decisions about evidential thresholds have given little attention to formal models capable of providing guidance on how precisely this ought to be done.

To appreciate the significance of this lacuna, suppose that one were convinced of the philosophical proposition that nonepistemic values should influence choices of inferential thresholds in science and wanted to know how to apply it in practice. Then several pressing questions immediately arise that cannot be addressed by informal discussions of inductive risk and that pose challenges to the traditional ideal of value-free science. We organize these into the following three groups: (1) *Prospective*: Where precisely should an inferential threshold be placed in a given context and why? (2) *Reverse Engineering*: What values are embedded in a given inferential threshold in a specific context? To what extent can those values be "reverse engineered" from the choice of threshold and details of the context, and how? (3) *Conceptual Clarification*: Can any of the ongoing disputes surrounding inferential thresholds be partially or wholly resolved by substituting a formal model for the prevailing informal modes of reasoning? Furthermore, are there any aspects of the received view, as reached through informal reasoning, that are due some reconsideration in the light of the greater generality of a formal model? This article draws on concepts from signal detection theory (SDT) to address these questions and illustrates this approach with several examples.

As explained in section 3, SDT provides a formal model in which questions about optimal choices of inferential thresholds can be answered for classification and hypothesis-testing problems. Therefore, it can provide answers to the three questions above. Given inputs about probabilities and costs, it can provide a rationale for deciding where to set an evidential threshold in a given context. Similarly, SDT can identify a range of assumptions about probabilities and costs that would lead to a given choice of threshold, which entails constraints on implicit valuations of costs if probabilities are known. Finally, the SDT framework can clarify concepts and identify omissions that may occur in reasoning about inferential thresholds, including some that have been discussed in literature on inductive risk and others that have not.

The main body of this article is organized as follows. In section 2, we briefly review philosophical literature on inductive risk, with an eye toward exposing gaps that arise as a result of the absence of a formal model. Section 3 presents the basic concepts of SDT and illustrates them with a schematic example of a diagnostic test. The subsequent three sections examine the three issues highlighted above. Section 4 discusses examples involving tuberculosis and lead poisoning to illustrate how evidential thresholds are determined within SDT, taking the misclassification costs estimated in the medical literature as given. In section 5, we examine the well-known 0.05 level of significance from the perspective of reverse engineering values inherent in this inferential threshold. The potential for conceptual clarification is ex-

https://doi.org/10.1086/702027 Published online by Cambridge University Press

plored in section 6 and is illustrated by base-rate neglect in the context of inductive risk. Finally, section 7 concludes with a discussion of the relevance of this framework for debates about inductive risk in the philosophy of science literature.

Before proceeding, we wish to emphasize two issues that we do not discuss here. First, we do not attempt to justify or defend the claim that nonepistemic values should influence scientists' choice of inferential thresholds.[1] Instead, our starting point is that if one accepts this claim—as many philosophers of science do—then a formal framework such as SDT is useful for mapping between values and inferential thresholds. Second, we do not attempt to answer questions about which nonepistemic values should be reflected in assessments of costs or how costs should be quantified. Although we give examples in which such quantification has been carried out, we do not inquire here into the propriety of how this was done.

2. The Argument from Inductive Risk. The classic presentation of the argument from inductive risk (AIR) is found in Rudner's (1953) accessible and compelling analysis. More recently, AIR has enjoyed a resurgence, in no small measure because of Douglas's (2000, 2009) influential work. Rudner's formulation of the argument begins with the premise that accepting and rejecting hypotheses is part of the daily work of scientists. Yet, the argument proceeds, scientific inferences from evidence to hypotheses are always more or less uncertain. Consequently, accepting a hypothesis requires a value judgment about how much uncertainty is tolerable in the situation. That value judgment, moreover, is often ethical in nature, since it depends on how bad, from a moral perspective, it would be to err in one direction or another. In Rudner's example, a higher inferential threshold would be required to accept that a toxic ingredient in a drug is not present at lethal levels than to accept that a batch of belt buckles stamped by a machine is not defective. Thus, according to Rudner, "How sure we need to be before we accept a hypothesis will depend on how serious a mistake would be" (1953, 2). And value judgments about the seriousness of mistakes will reflect the "moral standards" of those who make them (2). Rudner took the above reasoning to show that "the scientist as scientist does make value judgments" (2).

Douglas (2000, 2009) modifies and extends AIR in several respects. For example, rather than acceptance, Douglas speaks of asserting scientific claims, and rather than concluding that scientists actually do make value judgments, Douglas insists that they have a moral responsibility to do so (2009). In addition, Douglas (2000) extends AIR to encompass decisions involved in gen-

1. An anonymous referee points out that a number of epistemologists have recently argued that numerical loss functions can encode purely epistemic concerns (Joyce 1998; Pettigrew 2016).

https://doi.org/10.1086/702027 Published online by Cambridge University Press

erating evidence, including coding individual data points (e.g., as a malignant tumor or not) as well as selecting assumptions (e.g., threshold vs. linear dose response) that underlie statistical models used in data analysis.[2] However, despite these differences a recognizable pattern is evident. In general, variants of AIR insist, first, that inferences must be made from stochastic or otherwise uncertain evidence to one of a small number of discrete options, such as accept/reject/suspend judgment or assert/deny/refrain from commenting (see Steele 2012). And second, it is claimed that value judgments of an ethical or moral nature are relevant to deciding how to make such inferences. Classic objections to AIR challenge one of these two steps. For example, Jeffrey (1956) challenges the first by asserting that scientists should not accept or reject hypotheses and should instead limit themselves to reporting the probabilities of hypotheses given available evidence. In contrast, Levi (1960, 1962) agrees that accepting and rejecting hypotheses is a proper part of science but argues that in a scientific context these decisions should be driven exclusively by scientific values, such as explanatory scope and simplicity. A more recent criticism of AIR due to Betz (2013) can also be construed as targeting the second of the two steps just highlighted. That is, Betz claims that, if qualifications regarding uncertainty are thoroughly incorporated into the claims in question, then it is possible to avoid value judgments about how much certainty is enough.

In this article, our purpose is to explore implications of a formalization of AIR, with a focus on accepting or asserting hypotheses in light of statistical data. Consequently, we focus on aspects of the inductive risk literature that are relevant to the three questions highlighted above under the labels prospective, reverse engineering, and conceptual clarification. Since these questions all presume that one version or another of AIR is correct, we will not further review objections to the soundness of that argument here. Instead, we consider the somewhat less extensive subset of the inductive risk literature that is relevant to these three questions.

Let us begin with the prospective category, that is, questions about where to set an inferential threshold in a given context. As noted above, advocates of AIR propose no formal normative model of how probabilities and values should lead to decisions about inferential thresholds. Wald (1942) set out a general, abstract framework for determining the "best" statistical decision function and its associated critical region, which Churchman (1948) discusses in some detail. Although the post-Churchman AIR literature continues to acknowledge Wald, it has not, to our knowledge, developed any formal model of how probabilities and nonepistemic values should jointly determine inferential thresholds. As a result, it is unclear which inferential threshold

2. Some argue against expanding the scope of AIR in this manner (see Biddle and Kukla 2017).

https://doi.org/10.1086/702027 Published online by Cambridge University Press

should be chosen given background knowledge and values—even for one who wishes to follow AIR. Scarantino (2010), presenting an "illustration of one of the legitimate roles nonepistemic values can play in science" (422), argues that "the degree of confirmation required for accepting the HLA [Human Leucocyte Antigen] Hypothesis should be very high" (429). But how high, exactly?

Turn then to the second type of question: reverse engineering. It is sometimes claimed that influences of values are inherent or inevitable in choices of inferential thresholds (Wilholt 2009; Steele 2012; Winsberg 2012). However, several other philosophers have noted that inferential thresholds may be chosen for reasons unrelated to ethical or social values, such as convenience or inertia, and consequently that the motivations of a choice cannot be inferred from its ethically or socially significant impacts (Parker 2010; Morrison 2014; Steel 2016). In what sense, then, can values be said to be implicit in a choice of an inferential threshold? And how can one infer such implicit values from an inferential threshold plus constraints of the case? Can specific instances of scientific inference be 'reverse engineered' to reveal the values implicit in the standard of evidence used? We are not aware of any work in the inductive risk literature that provides a means for answering such questions.

Our final question type is conceptual clarification. The absence of a formal model can leave some important aspects of reasoning about inductive risk unclear. For example, Douglas (2009) argues that scientists have a moral obligation to consider harms that might result if claims they make are mistaken, but she does not similarly argue that they have an obligation to consider benefits that accrue if their claims are correct. This has led some to ask whether there is any reason to limit attention to harms of error in decisions about inferential thresholds or whether the benefits of getting it right should also be considered (Elliott 2011; Wilholt 2016). As we explain below, the SDT framework provides a simple and compelling answer to this question. In addition, discussions of inductive risk are typically framed in terms of probabilities of false negatives and false positives—that is, the chance of rejecting a claim when it is true or of accepting it when it is false. Such discussions make it easy to overlook the role of base rates in assessments of inductive risk, and indeed this role has received very little, if any, attention in the philosophical AIR literature. Nevertheless, base rates play an essential role in decisions about inferential thresholds in the SDT framework we describe.

3. Signal Detection Theory and Inductive Risk. In this section, we present the essential features of SDT and explain how it can be applied to inductive-risk problems. SDT is not the only formal framework within which the determination of inferential thresholds may be studied. For instance, Bayesian decision theory can be employed to solve for the action that maximizes

https://doi.org/10.1086/702027 Published online by Cambridge University Press

subjective expected utility, when the 'action' in question is adopting an inferential threshold (Berger 1985). Similarly, it would also be possible to follow an explicitly Bayesian reconstruction of reverse-engineering-mode SDT (see Kadane, Schervish, and Seidenfeld 1999). Nevertheless we elect to work within SDT—which can be viewed as a special case of Bayesian decision theory[3]—rather than full-blown Bayesian decision theory because SDT is a simplified, practitioner-friendly framework that is specialized for the very purpose of determining optimal inferential thresholds and that has been actively used for this purpose within the scientific community throughout the post–World War II period.

Preliminaries. SDT is a tractable and generic framework within which to analyze simple classification and hypothesis-testing problems. Examples of SDT may be found in diverse areas ranging from meteorology to medical diagnostics, as well as in quality control, credit scoring, and fraud detection in industry.[4] For concreteness, we introduce and illustrate the concepts of SDT in the context of simple two-state medical diagnosis, in which the two health-state categories are 'healthy' ($\neg D$) and 'disease present' (D). In the generic language of hypothesis testing, these correspond to the null hypothesis and the alternative hypothesis, respectively.

The diagnostic problem consists of inferring the patient's disease state $\{\neg D, D\}$ through observing the value of a *score variable* $X \in \mathbb{R}$. SDT provides a method for determining an optimal cutoff threshold $x^* \in \mathbb{R}$ above which it is inferred that the patient's health state is D, and the patient is said to test 'positive' for the disease. It is important for SDT that the score variable be real valued, as this is necessary for the construction of the receiver operating characteristics (ROC) curve, which we discuss below.

Some diagnostic tests are very simple, in that a patient's score-variable value is obtained from a single direct measurement—for instance, from measurement of the blood-sample concentration of a specific enzyme. In many diagnostic tests, however, the score-variable value is obtained by combining the results from several different direct measurements (of biomarkers) through a specifically developed (and validated) formula.

Sampling Distributions. Using the best available method—in some cases implemented through posthumous dissection—the true disease-D health state is determined for each subject in an N-strong representative sample of the population. This establishes the *base-rate prevalence* of the disease $p_D =$

3. We thank an anonymous referee for flagging these overlaps between SDT and Bayesian decision theory.

4. See, e.g., Swets, Dawes, and Monahan (2000) and Swets (2001) for a sample of diverse applications.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

n_D/N. The information also allows the construction of two *sampling distributions*, which we represent as the conditional distribution of score-variable values among the healthy $f(x|\neg D)$ and among those with the disease $f(x|D)$. Among the healthy there is variation in the score-variable values, and thus the empirical distribution has positive variance $\hat{\sigma}^2_{\neg D} > 0$, which is a measure of the *noise* present in the score variable. Similarly, there is variation in the score-variable values among those with the disease $\hat{\sigma}^2_D > 0$. If there were no overlap between the densities $f(x|\neg D)$ and $f(x|D)$, diagnosis would be trivial and, in principle, error free. The fact that the densities $f(x|\neg D)$ and $f(x|D)$ do overlap, almost invariably, sets the stage for explicitly striking a trade-off between errors of the first and second kind.

In practical diagnostic problems, SDT is applied directly to empirical sampling distributions. But for the purposes of developing the formal machinery of SDT, it is useful to work with families of known probability distributions. In this presentation, we use Gaussian sampling distributions, although in the literature one also finds applications of the chi-square, Poisson, gamma, power-law (exponential), geometric (Egan 1975), logistic, extreme-value (DeCarlo 1998), triangular, and beta (Marzban 2004) families of distributions.

Under the Gaussian assumption, we may specify the sampling distribution among the healthy as $X_{\neg D} \sim N(\mu_{\neg D}, \sigma^2_{\neg D})$ and among those with the disease as $X_D \sim N(\mu_D, \sigma^2_D)$. Crucially, diagnosis by applying a cutoff threshold can perform better than chance when $\mu_{\neg D} < \mu_D$.[5] Any given cutoff threshold x' simultaneously defines the true-negative rate ($\mathrm{TNR}_{x'} = 1 - \alpha_{x'}$) and the false-positive rate ($\mathrm{FPR}_{x'} = \alpha_{x'}$, i.e., type I error rate) as well as the true-positive rate ($\mathrm{TPR}_{x'} = 1 - \beta_{x'}$, i.e., statistical power) and the false-negative rate ($\mathrm{FNR}_{x'} = \beta_{x'}$, i.e., type II error rate; see fig. 1).

$$1 - \alpha_{x'} = P(X \leq x'|\neg D) = \int_{-\infty}^{x'} f(x|\neg D)\mathrm{d}x. \tag{1}$$

$$\alpha_{x'} = P(X > x'|\neg D) = \int_{x'}^{\infty} f(x|\neg D)\mathrm{d}x. \tag{2}$$

$$\beta_{x'} = P(X \leq x'|D) = \int_{-\infty}^{x'} f(x|D)\mathrm{d}x. \tag{3}$$

$$1 - \beta_{x'} = P(X > x'|D) = \int_{x'}^{\infty} f(x|D)\mathrm{d}x. \tag{4}$$

5. The opposite case $\mu_{\neg D} > \mu_D$ is not analyzed here, as it may be converted into the case we consider through a simple transformation.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

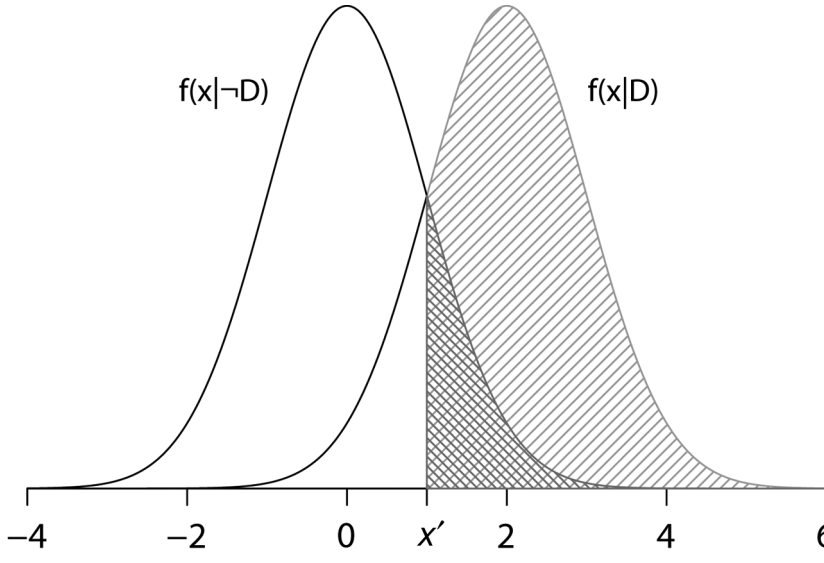


Figure 1. $\text{TPR}_1 = 1 - \beta_1 = 0.84$ (*45°* *hatch*) and the $\text{FPR}_1 = \alpha_1 = 0.16$ (*135°* *hatch*) generated by the cutoff threshold $x' = 1$ when $X_{\neg D} \sim N(0, 1)$ and $X_D \sim N(2, 1)$. Color version available as an online enhancement.

Receiver Operating Characteristics Curve. The extent to which diagnosis by using the cutoff threshold x' performs better than mere chance (i.e., in comparison with diagnosis by flipping a coin) depends on the extent to which the true-positive rate ($\text{TPR}_{x'} = 1 - \beta_{x'}$) exceeds the false-positive rate ($\text{FPR}_{x'} = \alpha_{x'}$). This of course varies parametrically with the cutoff threshold x'.

The ROC curve is the name given to this parametric plot of the locus of $(\alpha_{x'}, 1 - \beta_{x'})$ pairs generated by allowing x' to vary from $-\infty$ to ∞ (see fig. 2*b*). In so doing, the ROC curve encodes the highest-attainable statistical power $1 - \beta$ for any given fixed type I error rate α, holding constant the parameters of the sampling distributions.[6] In other words, the ROC curve encodes the best attainable combinations of TPR and FPR. Furthermore, because of the analytical ROC curve's smooth, differentiable form (unlike typical empirical ROC curves), it is particularly useful in capturing the entire range of possible trade-offs that may be struck between the TPR and the FPR.

The ROC curve is a routinely and widely used general performance measure for evaluating classifiers and diagnostic tests. The area under the curve (AUC) of an ROC curve is a summary measure of performance. ROC curves

6. The ROC curve permits the Neyman-Pearson lemma to be implemented directly by 'reading off' the greatest attainable statistical power associated with a fixed type I error rate (such as $\alpha = 0.05$).

https://doi.org/10.1086/702027 Published online by Cambridge University Press

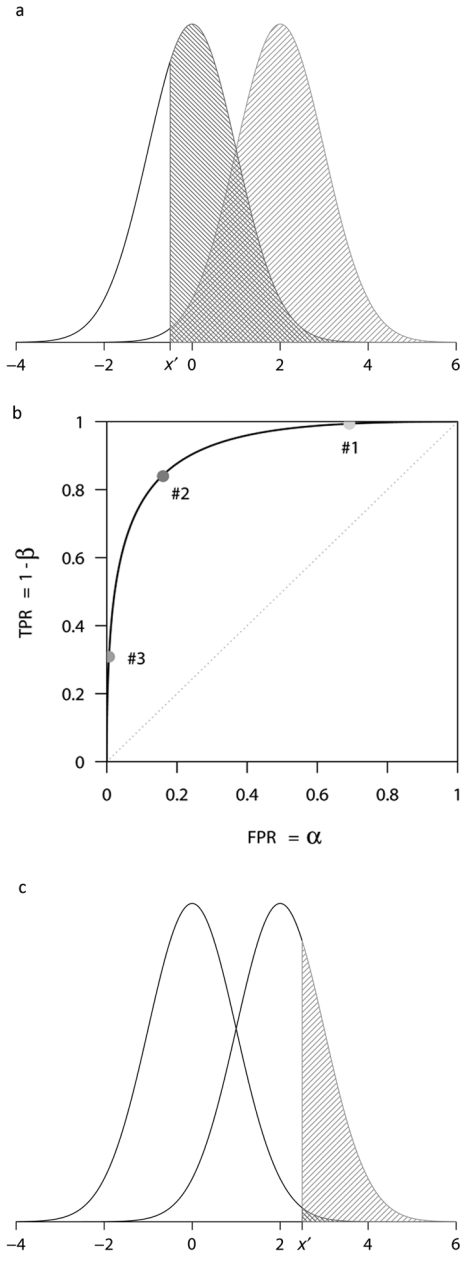


Figure 2. ROC curve for sampling distributions $X_{\neg D} \sim N(0, 1)$ and $X_D \sim N(2, 1)$. *a*), $x' = -0.5$, $\text{FPR}_{-.5} = 0.69$ (*135° hatch*), $\text{TPR}_{-.5} = 0.99$ (*45° hatch*). *b*), ROC curve. Markers indicate the (FPR, TPR) pairs associated with $x' = 2.5$ (*#3*), $x' = 1$ (*#2*), and $x' = -0.5$ (*#1*). *c*), $x' = 2.5$, $\text{FPR}_{2.5} = 0.01$ (*135° hatch*), $\text{TPR}_{2.5} = 0.31$ (*45° hatch*). Color version available as an online enhancement.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

of uninformative diagnostic tests coincide with the FPR = TPR diagonal and are characterized by AUC = 0.5 and performance no better than diagnosis by coin toss—regardless of the cutoff threshold. The ROC curve of a perfect classifier—assuming one could be found—would rise to the top left-hand corner where (FPR, TPR) = (1, 1), whereby AUC = 1. Diagnostic tests may be ranked by their AUCs, which in general fall within the open interval (0.5, 1).

As the ROC curve is a visualization of sampling-distribution integrals (see eqq.[2] and [4]), the AUC depends on the relationship between the sampling distributions. In our example of unit-equal-variance Gaussian sampling distributions, the AUC is determined uniquely by the normalized distance between the sampling distributions, which is called the *discriminability* of the diagnostic test $d' = (\mu_D - \mu_{\neg D})/\sigma$. Let $\Phi(\cdot)$ denote the standard-normal cumulative distribution function. The AUC in this example has a particularly simple representation: $\text{AUC} = \Phi(d'/\sqrt{2})$. Thus, in figure 2*b*, $\text{AUC} = \Phi(2/\sqrt{2}) = 0.92$, which is a fairly high value, given that the AUC may, under certain assumptions, be interpreted as the probability that a randomly selected positive case *i* will have a score value that is larger than a randomly selected negative case *j* (Green and Swets 1966).[7]

(Mis)classification Costs. In SDT, value judgments relating to outcomes are framed as costs, which in applications are generally expressed in (ratio-scale) monetary units.[8] But ultimately, it is the user of SDT who determines the units in which costs are denominated, as well as the types of factors that are deemed to be relevant in the diagnostic problem.[9] There clearly are technical, ethical, institutional, and sometimes legal questions to be confronted in determining the nature of the assay undertaken and the factors taken into consideration when computing each of the misclassification-cost-matrix elements. These considerations receive very little attention in the SDT literature, as determination of the numerical values with which to populate the misclassification-cost matrix is considered to be an extrastatistical, domain- and context-specific matter.

The misclassification-cost matrix comprises four elements, one for each state × inference combination. Each row in table 1 represents a state of na-

7. In the fig. 2*b* example, $P(x_i > x_j) = 0.92$.

8. The monetary-equivalent values of non-market-traded amenities—such as quality-adjusted life years (QALYs)—may be incorporated into the misclassification-cost matrix using, e.g., contingent-valuation methods, but this requires a separate, resource-intensive, methodologically challenging extrastatistical investigation of its own (Bobinac et al. 2014).

9. Hagen (1995), e.g., calculates misclassification costs in terms of changes to the mortality rate, excluding "outcomes such as inconvenience, pain, or monetary cost" (230).

https://doi.org/10.1086/702027 Published online by Cambridge University Press

TABLE 1. MISCLASSIFICATION-COST MATRIX

	Inference	
State	D	$\neg D$
D	C_{TP}	C_{FN}
$\neg D$	C_{FP}	C_{TN}

ture. The top row pertains to the disease-present state (D). In this state, the patient's score value is drawn from $f(x|D)$, and the diagnostic test delivers a true-positive result with associated cost C_{TP} when $x' < x$, else a false-negative result with associated cost C_{FN}. The bottom row pertains to the disease-absent state ($\neg D$). In this state, the patient's score value is drawn from $f(x|\neg D)$, and the diagnostic test delivers a false-positive result with associated cost C_{FP} when $x' < x$, else a true-negative result with associated cost C_{TN}.

Consider the misclassification-cost matrix in connection with an example involving a diagnostic test, where an effective treatment for the disease in question exists and is available to those tested. It is customary to set $C_{TN} = 0$. Under these circumstances, $C_{FN} > C_{TP} > 0$. That is, it is best not to have the disease at all, but if you have it, it is better to be correctly diagnosed and treated than not to be diagnosed. The costs in the bottom row in table 1 are those of not having the disease when the test result is positive or negative, respectively. In this case, $C_{FP} > C_{TN}$, since the false-positive test leads to harmful effects (e.g., anxiety, unnecessary medical procedures). Figure 3

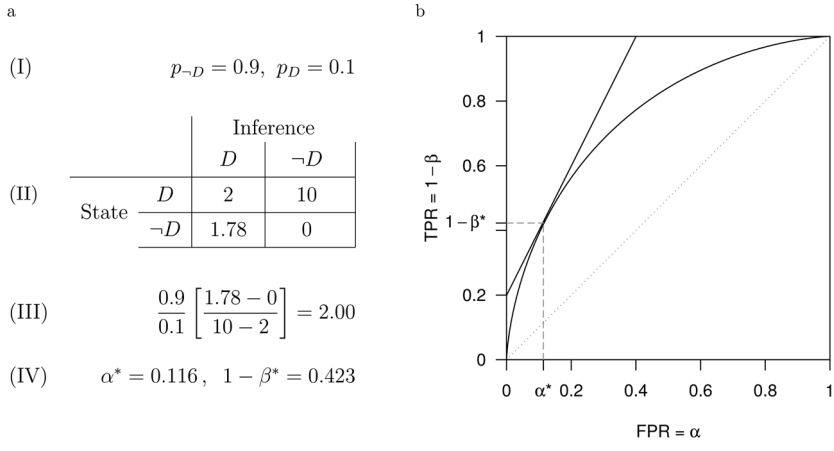


Fig. 3. Sampling distributions $X_{\neg D} \sim N(0, 1)$ and $X_D \sim N(1, 1)$. *a*, Parameters. *b*, ROC curve with tangency condition.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

provides a numerical example of what a classification matrix might look like in a medical diagnostic example such as this.

Note that the relationships among the misclassification costs will not always be as in figure 3. For example, false negatives are not always worse than false positives (i.e., it is not always the case that $C_{FN} > C_{FP}$). But we will assume in general that $C_{FP} > C_{TN}$ and similarly $C_{FN} > C_{TP}$. This assumption entails in effect that learning the truth about the hypothesis is better than coming to a false conclusion about it. Yet it is conceivable that this might not be the case and, for instance, that C_{TP} could exceed C_{FN}. This could happen in instances of harmful knowledge, such as Kitcher's example of "the imaginary discovery that vast quantities of energy could be released by mixing readily obtainable ingredients in just the right proportions, a discovery whose widespread publication would make our world an extraordinarily risky place" (2001, 149). In this example, falsely concluding that homemade weapons of mass destruction are not possible is better (less costly) than correctly discovering how to make them. However, in such an example the obvious implication is that the inquiry should not proceed, in which case questions about appropriate inference thresholds would be moot.

Loss Function. In principle, the optimal cutoff threshold x^* could be determined by optimizing any one of a large number of different potential 'loss functions'.[10] With few notable exceptions,[11] the loss function overwhelmingly employed in SDT is expected misclassification cost.[12] The mathematical expectation is taken with respect to the misclassification costs, where the probability weights are compounds between base-rate prevalences p_D, $p_{\neg D}$ and the conditional, x'-dependent probabilities $1 - \alpha_{x'}$, $\alpha_{x'}$, $\beta_{x'}$, and $1 - \beta_{x'}$. It is customary to incorporate also the fixed cost of implementing the diagnostic test, C_0. For concreteness, we present the expected-misclassification-cost expression $E(C)$ here in full:

$$\begin{aligned} E(C) &= C_{TP}P(\text{TP}) + C_{FN}P(\text{FN}) + C_{TN}P(\text{TN}) + C_{FP}P(\text{FP}) + C_0 \\ &= -[C_{FN} - C_{TP}] \cdot p_D \cdot (1 - \beta_{x'}) + [C_{FP} - C_{TN}] \cdot p_{\neg D} \cdot \alpha_{x'} \\ &\quad + C_{TN} \cdot p_{\neg D} + C_{FN} \cdot p_D + C_0 \,. \end{aligned} \tag{5}$$

10. In different fields, the loss function is also known as, inter alia, a 'goal function', a 'reward function', an 'objective function', or a 'penalty function'.

11. See Ulehla (1966) and Levi (1985) for non-risk-neutral expected utility or, with additional restrictions, Kaivanto (2014) for cumulative prospect theory.

12. In a Bayesian setting, one minimizes the risk function, which is the expected value of the loss function. Here we collapse terminology in the interest of simplicity and refer to the expected value of the misclassification costs as the 'loss function'.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

Optimality Condition. The optimal cutoff threshold x^* is obtained as the solution to the constrained minimization problem in which expected misclassification cost (5) is minimized subject to the constraint on $(\alpha_{x'}, 1 - \beta_{x'})$ given by the ROC curve. The slope of each iso-expected-cost contour[13]—and therefore also the slope of the cost-minimizing iso-expected-cost line at the optimal operating point on the ROC curve—is the ratio of expected incremental cost of misclassifying a healthy subject ($\neg D$) to the expected incremental cost of misclassifying a diseased subject (D).

Optimal Operating Point Condition:

$$\frac{p_{\neg D}}{p_D}\left[\frac{C_{\text{FP}} - C_{\text{TN}}}{C_{\text{FN}} - C_{\text{TP}}}\right] = \left(\frac{\text{d TPR}}{\text{d FPR}}\right)_{\bar{C}^*}. \tag{6}$$

Recall that $1 - \beta = \text{TPR}$ is represented on the vertical axis of the ROC space, while $\alpha = \text{FPR}$ is represented on the horizontal axis of the ROC space.[14] The term on the right-hand side of (6) is the slope of the tangent to the ROC curve at the expected-cost-minimizing point ($\bar{C}^*$). The term on the left-hand side is the above-mentioned ratio of expected incremental cost of misclassifying a healthy subject to the expected incremental cost of misclassifying a diseased subject. Since all of the terms on the left-hand side are known, we also know the slope of the tangent to the ROC curve at the expected-cost-minimizing point. The point along the ROC curve where equality (6) holds, which corresponds to a tangency condition between the cost-minimizing iso-expected-cost line and the ROC curve, determines the optimal test size and power $(\alpha^*, 1 - \beta^*)$ and, simultaneously, the optimal inferential threshold in the score variable x^*. Hence, the former may also be written as $(\alpha_{x^*}, 1 - \beta_{x^*})$.

In figure 3 we present a simple numerical illustration of how the optimal operating point condition (6) is implemented. Figure 3*a* presents (I) the base-rate probabilities, (II) the misclassification costs, and (III) the calculated slope of the iso-expected-cost line. Figure 3*b* presents the tangency condition graphically within the ROC space, which identifies the optimal test size

13. The loss function (5) defines a plane in the third dimension above the unit-square ROC space. This plane is canted down toward the top-left corner (0, 1). A *contour* of this plane—i.e., the set of all points in the plane that are at the same expected-cost 'elevation' $\bar{C}$—may be represented as a straight line within the two-dimensional ROC space. As all $(\alpha, 1 - \beta)$ points in this line are associated with the same expected misclassification cost $\bar{C}$, it is called an iso-expected-cost contour or an iso-expected-cost line.

14. The notation $(\text{d}x/\text{d}y)_{\bar{z}}$ denotes the *derivative* of x with respect to y at the point where $z = \bar{z}$ holds.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

and statistical power as $(\alpha^*, 1 - \beta^*) = (0.116, 0.423)$. It may be helpful to consider what sort of situation could correspond to this numerical example. Since C_{FN} is by far the greatest cost, the disease is one that has very severe effects if left untreated. Moreover, note that in the example C_{FP} is almost equal to C_{TP}. In other words, being incorrectly diagnosed as having the disease is almost as bad as being diagnosed and actually having it. Such an assessment of costs would be sensible in a case in which diagnosis leads directly to treatment, for instance, without further, more accurate, diagnostic tests, and the treatment is very effective.

The larger the ratio in condition (6), (i) the farther to the left the optimal operating point occurs on the ROC curve, (ii) the smaller the optimal test size α^* and power $1 - \beta^*$, and (iii) the larger the cutoff-threshold score x^*. These properties i–iii are general and independent of the specific parametric form of the sampling distributions. In connection with our disease example, this means that if the disease is very rare and if a mistaken positive diagnosis is more harmful than a mistaken negative diagnosis, then a very strict evidential threshold should be required for accepting that the disease is present (i.e., for rejecting the null hypothesis $\neg D$). Conversely, a less strict threshold would be appropriate if the disease is common or if mistakenly thinking the disease is absent is much worse than mistakenly thinking it is present. Our numerical example illustrates the latter situation. That is, it is an example in which false negatives are much worse than false positives, which is reflected in the optimal test size being larger than the conventional 0.05. The optimal test size would be larger still but for the low prevalence rate of $p_D = 0.1$. Note that the difference between C_{FN} and C_{TP} is important for determining how harmful it is, in expectation, to mistakenly conclude the disease is absent. For instance, if there is no effective treatment for the disease, then little may be lost by not diagnosing it. In such a case, C_{TP} would be close to C_{FN}, and consequently the denominator of (6) would be small even if C_{FN} were very high.

In the medical SDT literature, the bracketed term in condition (6) is commonly referred to as the cost-benefit (C/B) ratio. Since negative costs are benefits just as negative benefits are costs—remember that SDT requires costs to be ratio-scale measures—the medical SDT literature uses the following equivalent form of the bracketed misclassification-cost-difference term

$$\left[\frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}}\right] = \left[\frac{C_{FP} - C_{TN}}{-(C_{TP} - C_{FN})}\right] = \left[\frac{C_{FP} - C_{TN}}{B_{TP} - B_{FN}}\right]. \tag{7}$$

In what follows, we discuss how the SDT framework can usefully address the three questions we highlighted in the introduction under the labels prospective, reverse engineering, and conceptual clarification.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

4. Prospective Examples. Prospective questions have to do with exactly where to set an evidential threshold in a specific context, given information about costs and probabilities. Although it is relatively rare for scientists to use SDT for this purpose, some applications exist. For example, Cantor et al. (1999) performed a structured survey of the medical literature between 1976 and 1997 inclusive. Their search identified 48 articles explicitly mentioning a C/B ratio or an explicit method for determining the cutoff threshold. Altogether, 14 articles included a C/B ratio as part of the ROC analysis. In this section, we discuss two examples: tuberculosis and lead poisoning. In addition to illustrating how SDT can work in practice to determine inferential thresholds, these two examples illustrate philosophically significant themes about inductive risk that will be further elaborated in subsequent sections of this article. The tuberculosis example illustrates the importance of base rates, while the lead poisoning example illustrates misclassification-cost matrix operationalization and the effect of improvements in scientific discriminability (d' and AUC) on the inferential threshold.

Tuberculosis. The $1/400 = 0.0025$ C/B ratio reported by Lusted (1971) pertained to the US population in 1971, and of course this would be different if one were to reconsider the costs specific to particular subpopulations, whether they would be identified by demographic variables or by geography. Lusted (1971, 419) arrives at this C/B ratio by eliciting the physician's attitudes concerning the relative value of true-positive versus true-negative diagnoses, followed by "questions about the relative cost of diagnostic errors compared with the value of correct diagnoses," in each case disaggregating across numerous subjective-value categories (1, value to my self-esteem of correct diagnosis; 2, value to referring physicians of correct diagnosis; 3, value to my reputation with referring physicians; 4, value to patient of correct diagnosis; 5, value to my reputation with patients; 6, value to society of correct diagnosis, etc.) and objective-value categories (1, my professional fee, etc.). Thus, Lusted's (1971) misclassification costs reflect nonepistemic values as this term is usually understood by philosophers. That is, the values in question are not focused on truth, explanatory power, or other knowledge-seeking aims commonly associated with science but instead concern the human welfare and economic impacts of correct and incorrect diagnoses. Consequently, while the sharpness of the distinction between epistemic and nonepistemic values may be questioned, we take the tuberculosis example discussed in this section to illustrate AIR.

With a known C/B ratio, the slope of the iso-expected-cost line then depends on the base-rate odds of the healthy relative to those with the disease $p_{\neg D}/p_D$. A prevalence of 0.0005 implies an iso-expected-cost line slope of $0.0025 \times (0.9995/0.0005) = 5.00$. If, however, the prevalence

https://doi.org/10.1086/702027 Published online by Cambridge University Press

is $117.8/100,000 = 0.001178$, as in Romania in 2006, then the slope of the threshold-determining iso-expected-cost line is $0.0025 \times (0.9988/0.001178) = 2.12$. The prevalence of 0.001178 is of course an average that overstates the prevalence for most of the Romanian population, as most cases of tuberculosis are concentrated within particular subpopulations. And if the prevalence is $870/100,000 = 0.0087$, as in Lusaka province of Zambia in 2009 (Ayles et al. 2009), the slope of the iso-expected-cost line falls to $0.0025 \times (0.9913/0.0087) = 0.28$. Of course these calculations presume that the C/B ratio applicable in Romania and Lusaka is the same as that which applies in the United States—an assumption made here to retain comparability between the optimality conditions as the base rate varies from one population to another. The associated optimal operating points are illustrated in figure 4, which assumes that the diagnostic test has unit discriminability $d' = 1$. The greater the iso-expected-cost line's slope, the smaller the α^* and $1 - \beta^*$. The diagnostic test's optimal operating point is (0.0159, 0.1362) for the US population, (0.103, 0.4005) for the Romanian population, and (0.78, 0.9637) for the Zambian population. The comparatively high prevalence of tuberculosis in Zambia entails a liberal cutoff threshold, whereas the comparatively low prevalence of tuberculosis in the United States entails a more conservative cutoff threshold.

Lead Poisoning. Note that the bracketed cost term in condition (6) may not be reduced to a single number without explicitly identifying each of the costs (C_{TN}, C_{FP}, C_{TP}, C_{FN}) separately. DeBaun and Sox (1991) trace the direct medical costs and the indirect human-capital-based costs associated with each (mis)classification category of lead poisoning. As in the tuberculosis example just discussed, these monetary-value-denominated costs unambiguously reflect nonepistemic values. Table 2 shows that the indirect costs, especially those associated with false-negative diagnoses, far exceed the other categories of misclassification cost.

Combined with the 12% nationwide prevalence of lead poisoning at the time (DeBaun and Sox 1991, 127), the iso-expected-cost line's slope is therefore

$$\left(\frac{1 - 0.12}{0.12}\right)\left[\frac{168 - 63}{6,159 - 4,361}\right] = \left(\frac{0.88}{0.12}\right)\left[\frac{105}{1,798}\right] \tag{8}$$
$$= (7.\dot{3})[0.0584] = 0.428\,.$$

The optimal operating point depends not only on the slope of the iso-expected-cost line (8) but also on the AUC, which in turn is increasing in the normalized distance between the sampling distributions d'. Figure 5 shows the relationship

https://doi.org/10.1086/702027 Published online by Cambridge University Press

a

Country	Base rate	Slope	α^*	$1-\beta^*$
US	0.000500	**5.00**	0.0159	0.1362
Romania	0.001178	**2.12**	0.1030	0.4005
Zambia	0.008700	**0.28**	0.7800	0.9637

b

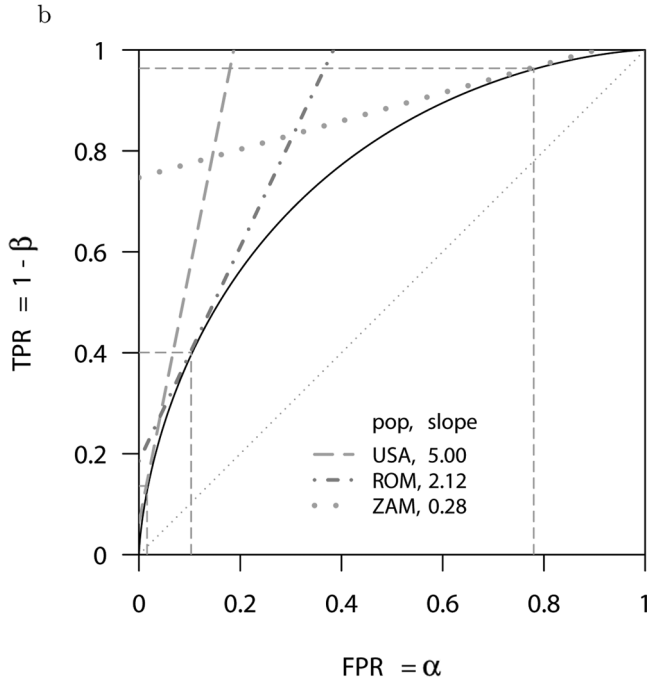


Figure 4. Optimal operating point determination for tuberculosis testing in the United States, Romania, and Zambia using sampling distributions $X_{\neg D} \sim N(0, 1)$ and $X_D \sim N(1, 1)$ and $C/B = 1/400$. *a*, Parameters. *b*, ROC curve with tangency conditions. Color version available as an online enhancement.

between the optimal FPR α^* obtained with the iso-expected-cost line slope (8) as the AUC increases from 0.5 to 1. For very low-discriminability AUCs where the ROC curve virtually coincides with the principal diagonal, (8) entails that the optimal cutoff threshold x^* is far in the left tail, and therefore $\alpha^* = \int_{x^*}^{\infty} f(x|\neg D)\mathrm{d}x \approx 1$. As discriminability and the AUC increase, the optimal cutoff threshold x^* moves to the right out of the left tail, and α^* begins to fall. However, given that the iso-expected-cost line's slope is very shallow

https://doi.org/10.1086/702027 Published online by Cambridge University Press

TABLE 2. Misclassification Costs (US$)

	Direct Cost	Indirect Cost	Total
C_{TP}	1,463	2,898	4,361
C_{TN}	63	0	63
C_{FP}	168	0	168
C_{FN}	63	6,096	6,159

(0.428), α^* begins to approach the conventional $\alpha = 0.05$ level only when the AUC approaches 1.

5. Reverse-Engineering Example

Overview. In this section, we consider the relevance of SDT to the notion that social or ethical values can be implicit in choices of inferential thresholds

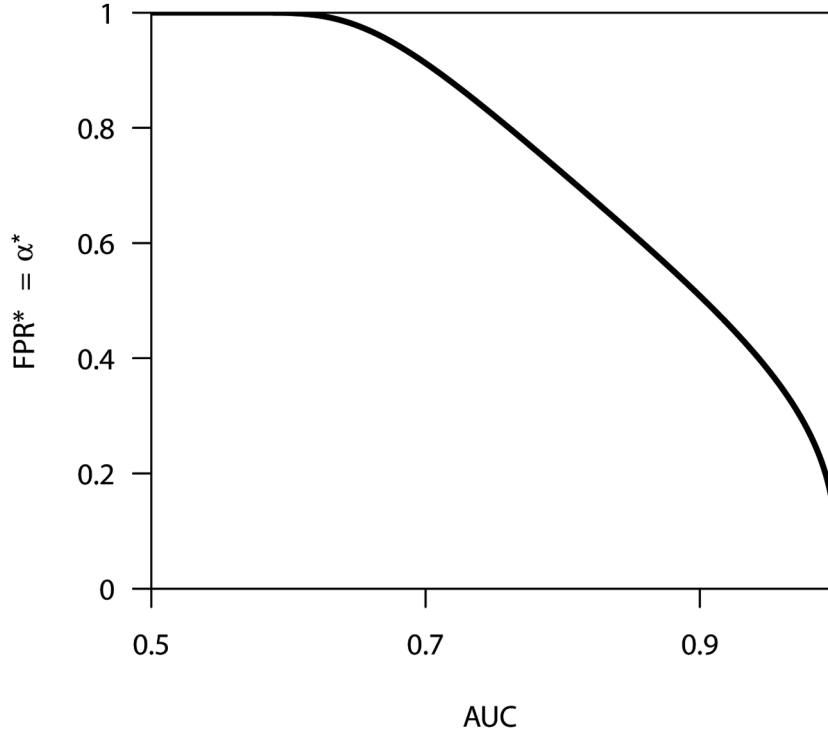


Figure 5. Lead-poisoning test's α^* computed for all possible values of AUC $\in$ [0.5, 1], holding misclassification costs and the base rate constant.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

even when such values are not explicitly invoked as reasons for the choice. We approach this question by considering the case of the widespread convention of 0.05 as the cutoff for statistical significance. We suggest that the notion of values implicit in a choice of inferential threshold be construed in a rational-reconstruction sense rather than in a psychological sense. That is, implicit values are those that, if accepted, would support a choice of inferential threshold but are not necessarily the values that actually motivate practitioners to adopt it. Thus, we consider what those values might be in the case of the $\alpha = 0.05$ convention, bearing in mind that according to SDT the answer to this question also depends on the base rate and the area under the ROC curve.[15]

Embedded Values in the $\alpha = 0.05$ Convention. Under null hypothesis significance testing, many fields have adopted an inferential-threshold convention that is applied as a de facto requirement for publication. Specifically, many fields have adopted $\alpha = 0.05$ as their inferential-threshold convention, although some fields have opted for stricter thresholds.[16] Insofar as a scientific field is characterized by the application of a particular collection of empirical methods to study a particular collection of questions and hypotheses, it may in some sense be reasonable to assume 'roughly similar' misclassification costs and 'roughly similar' base rates. However, the mathematical structure of SDT reveals that rigid application of a particular evidential threshold presupposes a precise nonlinear relationship between misclassification costs and base rates—which furthermore varies with discriminability (i.e., AUC).

In order for $\alpha = 0.05$ to be optimal, the product of the base-rate odds $(p_{\neg D}/P_D)$ with the incremental-misclassification-cost ratio $[(C_{\text{FP}} - C_{\text{TN}})/(C_{\text{FN}} - C_{\text{TP}})] = [IC_{\text{FP}}/IC_{\text{FN}}]$ must be equal to the slope of the tangent to the ROC curve at $\alpha = 0.05$:

$$\left(\frac{p_{\neg D}}{p_D}\right)\left[\frac{IC_{\text{FP}}}{IC_{\text{FN}}}\right] = \left(\frac{\text{d FPR}}{\text{d TPR}}\right)_{\alpha=0.05} \tag{9}$$

For our Gaussian sampling distributions, the slope of the tangent to the ROC curve with $d' = 2.0$ and AUC $= 0.92$ at $\alpha = 0.05$ is 3.63. Therefore, only those combinations of $(p_{\neg D}/P_D)$ and $[IC_{\text{FP}}/IC_{\text{FN}}]$ whose product is 3.63 are consistent—in the expected misclassification-cost-minimizing sense—with

15. Throughout this section we maintain the assumption that sampling distributions are equal-variance Gaussian. Qualitatively similar results would be obtained with other symmetric, equal-variance sampling distributions.

16. Particle physics employs a 5σ ($p < .0000003$) threshold. This reflects standard good statistical practice of adjusting the inferential-threshold downward whenever multiple simultaneous hypothesis tests are performed.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

$\alpha = 0.05$. The dashed curve in figure 6 represents all such combinations. Meanwhile, the slope of the tangent to the ROC curve with $d' = 0.2$ and AUC $= 0.56$ at $\alpha = 0.05$ is 1.36. The dotted curve in figure 6 represents all those combination of $(p_{\neg D}/P_D)$ and $[IC_{\text{FP}}/IC_{\text{FN}}]$ whose product is equal to 1.36 and therefore consistent with $\alpha = 0.05$ for $d' = 0.2$ and AUC $=0.56$.

When discriminability is $d' = 2.0$ and AUC $= 0.92$, all hypotheses featuring base-rate odds and incremental-misclassification-cost ratios that fall above the dashed curve receive biased treatment in that the $\alpha = 0.05$ inferential threshold is suboptimally liberal in the sense of making it too easy to reject the null hypothesis. Conversely, all hypotheses whose base-rate odds and incremental-misclassification-cost ratios fall below the dashed curve receive biased treatment in that the $\alpha = 0.05$ inferential threshold is suboptimally conservative in the sense of making it too difficult to reject the null hypothesis.

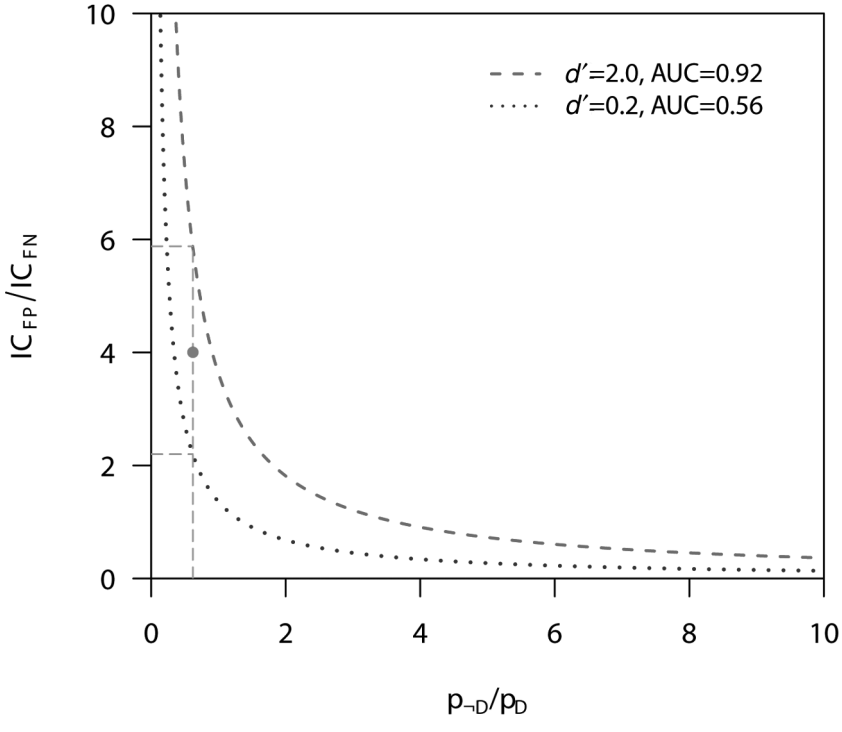


Figure 6. Schedules between base-rate odds and incremental-misclassification-cost ratios under which the optimal false-positive rate is held constant at $\alpha = 0.05$. Color version available as an online enhancement.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

Of course, given the costs and inherent epistemic difficulties involved in estimating base-rate and incremental-misclassification-cost ratios, it would be unreasonable to demand that every scientific hypothesis test be implemented with its own individually optimized inferential threshold α^*. In addition to these general concerns, applying SDT to scientific hypotheses raises special difficulties in assessing misclassification-cost ratios. The practical applications of the hypothesis, if any, may be difficult for researchers to anticipate, as may be its impact on science. Moreover, a hypothesis may be tested multiple times, and misclassification costs might not be constant across the several tests. For instance, one more experiment to test a hypothesis that has already been the subject of extensive research may have less impact than a pathbreaking study that tests a hypothesis for the first time.[17] Given such considerations, justifications of the $\alpha = 0.05$ convention—or indeed any other conventional cutoff—are best interpreted as attempts to show that the cutoff is a good-enough approximation in a particular field (Wilholt 2009). Let us assume, then, that the choice of $\alpha = 0.05$ in a scientific field should be justified in this manner. Then we can ask what sorts of general assumptions about base rates and misclassification costs are presumed by such reasoning.

The pattern revealed in figure 6 is that $\alpha = 0.05$ entails either that IC_{FP} greatly exceeds IC_{FN} or $p_{\neg D}$ greatly exceeds p_D—but not both as this would entail an even lower α. For example, in the $d' = 2.0$ and AUC $= 0.92$ line, if the base-rate odds are 2, then the incremental-misclassification-cost ratio must be slightly less than 2. If the base-rate odds were 2 while the incremental-misclassification-cost ratio were, say, 5, then $\alpha = 0.05$ would be suboptimal insofar as permitting too many type I errors. And if the base-rate odds are 10, then $\alpha = 0.05$ entails that the incremental-misclassification-cost ratio is less than 1 (i.e., IC_{FN} is greater than IC_{FP}. Conversely, if the incremental-misclassification-cost ratio is 10, then the base-rate odds are below 1 (i.e., $p_D > p_{\neg D}$). We consider the implications of this pattern by examining two types of case in which the $\alpha = 0.05$ convention might be challenged.

Begin with what one might call a Popperian outlook on science, according to which the vast majority of hypotheses are false and priority should be placed on avoiding false positives. In other words, under this assumption, $p_{\neg D}$ far exceeds p_D (e.g., by a factor of at least 6 to 1), while IC_{FP} is significantly greater than IC_{FN} (e.g., at least two times greater). In this case, inspection of figure 6 suggests that the $\alpha = 0.05$ convention is too lenient and that a smaller value of α, and hence a stricter standard for rejecting null hypotheses, would be called for. Adopting the $\alpha = 0.05$ convention in such a context, then, would be problematic, since it might contribute an unacceptable rate of errors or failures of replication. To adopt the $\alpha = 0.05$ con-

17. Note that the practice of fixed-level testing has been shown to be incoherent in the Bayesian setting (see, e.g., Schervish, Seidenfeld, and Kadane 2002).

https://doi.org/10.1086/702027 Published online by Cambridge University Press

vention in a scientific field, then, is to implicitly assume that the situation just described does not obtain there. In other words, it is to assume that either the base rate of true hypotheses is reasonably high or that false positives are not much worse than false negatives. It is unclear to what extent such assumptions are correct in areas of science in which the $\alpha = 0.05$ convention is prevalent. However, SDT shows that any attempt to justify the $\alpha = 0.05$ convention must take such matters into consideration.

Next consider a case in which false negatives are assumed to be far worse than false positives, say, by a factor of 10. Some philosophers have argued that, in such cases, the $\alpha = 0.05$ convention is unjustified and that a higher value of α should be chosen instead (Shrader-Frechette 1991; Cranor 1993). However, the above analysis shows that such arguments are crucially incomplete. Even if IC_{FN} exceeds IC_{FP} by a factor of 10, $\alpha = 0.05$ or an even lower setting of α may still be justified if the hypotheses tested in the field are overwhelmingly false.

We draw the following conclusions from the discussion in this section. First, the SDT framework can be used to identify sets of assumptions about base rates and incremental misclassification costs that, given the ROC curve associated with a test, could be used to rationalize an antecedently made choice of α. In this sense, SDT can be said to provide a basis for reverse engineering values that are inherent or implicit in choices of evidential thresholds in a context. Second, the SDT framework provides a basis for critically examining arguments made with respect to the $\alpha = 0.05$ convention, either that it is supported by a Popperian view of science or that it is inappropriate when false negatives are worse than false positives. As such, we hope that the framework we propose here can usefully contribute to discussions of when and where the $\alpha = 0.05$ convention is—and is not—reasonable.

6. Conceptual Clarification. In this section, we discuss three conceptual clarifications that flow from the SDT framework proposed here. The first of these, base-rate neglect, has already featured in several examples discussed in preceding sections. The second concerns whether only harms resulting from errors should be considered in setting inferential thresholds or whether benefits ensuing from correct inferences should also be considered. Finally, we consider the idea that, as scientific certainty increases, the relevance of values to choice of inferential thresholds diminishes. The effect of base rates is a topic that has been largely neglected in the AIR literature. The other two issues have been discussed in philosophical literature, but the SDT framework nevertheless provides valuable clarifications and refinements.[18]

18. Although these sections continue to rely on the equal-variance Gaussian sampling distributions machinery, qualitatively comparable results would be obtained with other symmetric, equal-variance sampling distributions.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

Base-Rate Neglect. The role of base rates has not featured prominently in the AIR literature. Bayesian prior probabilities have not featured prominently either. For instance, Douglas only mentions prior probabilities in a footnote commenting on Bayesian statistics, rather than in connection with the determination of inferential thresholds: "The Bayesian framework, has yet to be shown useful in real world contexts where both likelihoods and priors are disputed. . . . When both likelihoods and priors are disputed, abundant evidence may still never produce a convergence of probability" (2009, 183 n. 15). It is clear that base rates feature in the optimal operating point condition (6). It is also clear that in many types of real-world inferential problems—such as in diagnosing a disease or in detecting faulty products on a manufacturing line—base rates are indeed expressible as ratios of frequencies. In the tuberculosis example, we illustrated the large effect that base rates can have on the location of the optimal inferential threshold. Increasing the base rate from 0.0005 to 0.0087 shifted the optimal inferential threshold (α^*) by a factor of 49 from 0.0159 to 0.78. And in section 5 we illustrated the three-way interaction between base rates, incremental-misclassification costs, and discriminability (AUC). Hence, it is not possible to discuss difficulties related to a rigid application of the $\alpha = 0.05$ criterion without making an assumption about the base rate.

SDT shows that both base-rate odds and discriminability (i.e., AUC) are moderators of the effect of changes in misclassification costs on the optimal inferential threshold. Nevertheless base-rate odds are a much more powerful 'lever' with which to influence the optimal inferential threshold than misclassification costs. This finding has been noted and replicated in empirical studies (Wolfe, Horowitz, and Kenner 2005; Wolfe et al. 2007; Evans, Birdwell, and Wolfe 2013). But the analytical structure of SDT also reveals why this is the case. On a basic level it is partly due to the fact that any change in the base rate has two separate impacts on the prior odds term,[19] whereas there is no comparable 'double impact' of changes in misclassification costs on the incremental-misclassification-cost ratio. On a deeper level, the SDT model reveals that the inferential threshold in the underlying score variable x^* is highly sensitive to—indeed completely dominated by—small base rates, when they are present. If either base rate becomes very small, the inferential threshold moves far into the sampling distribution's tail, accommodating the improbable state.[20] For equal-unit-variance Gaussian sampling distributions with $\mu_{\neg D} = 0$, the optimal cutoff threshold in the score variable may be obtained directly from

19. Any small increase in p_D entails a corresponding decrease in $p_{\neg D} = 1 - p_D$. Hence, the ratio $p_{\neg D}/p_D$ decreases not only because the denominator p_D increases but also because of the corresponding decrease in the numerator $p_{\neg D}$.

20. Below it is shown that when $p_{\neg D} \to 0$, $x^* \to -\infty$, while when $p_D \to 0$, $x^* \to \infty$.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

$$x^* = \frac{1}{\mu_D}\left(\ln(C_{\text{FP}} - C_{\text{TN}}) - \ln(C_{\text{FN}} - C_{\text{TP}}) + \ln(p_{\neg D}) - \ln(p_D) + \frac{\mu_D^2}{2}\right). \quad (10)$$

Notice that as $p_{\neg D} \to 0$, $\log(p_{\neg D}) \to -\infty$, and thus $x^* \to -\infty$. Alternatively if $p_D \to 0$, then $-\log(p_D) \to \infty$, and thus $x^* \to \infty$.

The necessity of incorporating base rates into the analysis of inferential thresholds is among the key implications of our formal modeling approach. This is not to suggest that existing analyses are rendered logically invalid. Abstracting from other factors, increasing the cost of type I errors continues to suggest a more conservative inferential threshold, while increasing the cost of type II errors continues to suggest a more liberal inferential threshold. However, the base rate is an exceptionally strong modifier of the effect of error costs on the inferential threshold. The effect of any finite change in error costs becomes arbitrarily small if either base rate ($p_{\neg D}$ or p_D) is sufficiently small. This means that under certain base-rate configurations, the impact of changes in error costs is rendered inconsequential.

In addition, variation in base rates among subpopulations is often extremely important for reverse engineering value judgments that are implicit in a choice of inferential threshold. The prevalence of tuberculosis, for instance, is higher in identifiable subpopulations. Fournet et al. (2006) report that in 2003 the incidence of tuberculosis within Rio de Janeiro state prisons was 15 times higher than in the general state population. A more recent study estimates the incidence rate among prisoners to be more than 20 times that in the general Brazilian population (Carbone et al. 2015). In the US prison population, the prevalence of tuberculosis can be up to 17 times that in the general US population (Roberts et al. 2006). And in the low-incidence country of the Netherlands, 61% of registered tuberculosis patients were born abroad, even though the foreign-born population comprises only 11% of the overall population (Kik et al. 2009). In each of these cases, use of the general-population inferential threshold for the purpose of testing the subpopulation can be interpreted as placing a lower value on the lives of members of the high-base-rate subpopulation.

Further Remarks. The SDT framework suggests solutions for an array of issues, including some that are currently being debated within philosophy. Here we briefly note two of these ongoing debates along with the reframing suggested by SDT, deferring full and formal treatment to future work.

Harms Alone versus Harms and Benefits. Among philosophers who accept the general premise that scientists have a moral obligation to consider the potential nonepistemic consequences of mistakenly accepting or rejecting a hypothesis, consensus is yet to crystallize over whether only harms alone should influence the inferential threshold or whether the associated

https://doi.org/10.1086/702027 Published online by Cambridge University Press

benefits should also be taken into account (Douglas 2009; Elliott 2011; Wilholt 2016). For Douglas (2009) the very reason why scientists should incorporate nonepistemic values into their reasoning about inferential thresholds is that scientists bear a moral responsibility for the unintended harms flowing from their actions due to recklessness or negligence (68–71). Hence, it is the potential harms flowing from errors in scientific inference that scientists are morally required to incorporate into the determination of inferential thresholds. Some philosophers have asked whether "the potential *benefits* of accepting or rejecting claims erroneously" should also be considered or indeed whether also the benefits and harms of correctly accepting or rejecting claims should also be considered (Elliott 2011, 315). Wilholt (2016) argues that responsible epistemic risk management cannot exclude the nonepistemic consequences (benefits and harms) of correct classification and that imposing an asymmetry between the consequences of erroneous and correct classification cannot be justified.

SDT's expected-misclassification-cost loss function places harms and benefits on an equal footing. It also places the consequences ensuing from misclassification on an equal footing with the consequences ensuing from correct classification. SDT does this not for ethical reasons but for mathematical reasons and for reasons inherent in the expected-misclassification-cost loss function. Benefits are equivalent to negative costs; that is, a benefit is created by reducing cost. Costs are equivalent to negative benefits; that is, a cost is incurred when benefits are reduced. This relationship underpins the practice in the medical SDT literature of referring to the 'cost-benefit ratio' as derived above in equation (7). Furthermore, there is also a fundamental sense in which cost calculus inherently requires both costs and benefits to be fully accounted for, in that the concept of 'opportunity cost' is defined as 'forgone benefit'. Without accounting for benefits, it is not possible to define opportunity costs.

Thus, in examples involving diagnosis, the harm of a false negative often depends on an opportunity for effective treatment being missed. That is why it is sensible for the denominator in (6), the optimal operating point condition, to include the difference between C_{FN} and C_{TP} rather than C_{FN} alone. When effective treatment exists and is accessible, C_{TP} is less than C_{FN}, and hence harm is done by a missed diagnosis. The relevance of both C_{FP} and C_{TN} for assessing the harms of false positives is somewhat less apparent, given the tendency to treat true negatives as the default state. However, C_{TN} can also be viewed as conferring benefits, whose presence are relevant to the extent to which C_{FP} exceeds C_{TN}. For example, if the patient is deeply anxious before the test because of believing that the disease is present, then a true negative result confers a substantial emotional benefit. In such a case, a false positive is harmful in part because it entails forgoing the benefit of correcting an anxiety-inducing false belief.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

Do Values Become 'Less Important' as Uncertainty Decreases? The answer turns on what is meant by 'decrease in uncertainty' and 'become less important' respectively. Within the SDT framework it is natural to operationalize decreases in uncertainty as increases in the resolving power of the scientific experiment (i.e., as increases in d' and AUC). But as we show below, decreases in such scientific uncertainty cause nonmonotonic changes in either α^* or $1 - \beta^*$. In other words, values continue to drive nonobvious changes in α^* and $1 - \beta^*$ as scientific uncertainty decreases. The sense in which values become 'less important' is more subtle and is induced not by the mere change in d' and AUC magnitude but by the associated *increase in the curvature* of the ROC curve as d' and AUC increase. For any given percentage change in $(p_{\neg D}/P_D)[IC_{\text{FP}}/IC_{\text{FN}}]$, the relative response in (β^*/α^*) becomes smaller as scientific uncertainty decreases. The sensitivity of (β^*/α^*) to changes in values is attenuated as uncertainty decreases.

Consider these two points in turn. The ability of the experiment to distinguish between $H_0 : \neg D$ and $H_1 : D$ is summarized by $d' = (\mu_D - \mu_{\neg D})/\sigma$. The larger d', the larger the AUC, and if the FPR $= \alpha$ is held constant, then the remaining uncertainty associated with the inference (i.e., FNR $= \beta$) is necessarily smaller. However, the premise within the AIR literature is that fixed inferential thresholds such as $\alpha = 0.05$ are inappropriate insofar as they do not reflect error-cost considerations, nor, as we have shown, do they reflect base-rate considerations. Taking these considerations seriously, we see that the inferential threshold changes as d' and the AUC increase. Figure 7 shows the paths traced out by the optimal operating points within the ROC space as AUC varies within the interval (0.5, 1), for several iso-expected-cost-line slopes. These constant-slope loci are known as *isoclines*. Figure 7 exhibits isoclines for five different iso-expected-cost-line slopes, two of which are less than 1, and two of which are greater than 1. The $(p_{\neg D}/P_D)[IC_{\text{FP}}/IC_{\text{FN}}] = 1$ case is straightforward and conforms with lay intuition, in that both α^* and β^* decrease monotonically as AUC increases. However, for all other slope values $(p_{\neg D}/P_D)[IC_{\text{FP}}/IC_{\text{FN}}] \gtrless 1$ the relationship is nonmonotonic for either α^* or β^*.

Thus, as the resolving power of science improves, we cannot in general infer a monotonic reduction in inferential-error probabilities when scientists are appropriately incorporating nonepistemic consequences into their inferential thresholds. As the AUC increases from very low levels, the likelihood of a type I error increases if the expected incremental cost of misclassifying a negative exceeds the expected incremental cost of misclassifying a positive. Conversely, if the expected incremental cost of a false positive exceed that of false a negative, then the likelihood of a type II error rises as AUC increases from very low levels. Value considerations—in the form of the slope of the iso-expected-cost line—determine the critical level of AUC above which further increases in the AUC reduce the likelihood of erroneous results

https://doi.org/10.1086/702027 Published online by Cambridge University Press

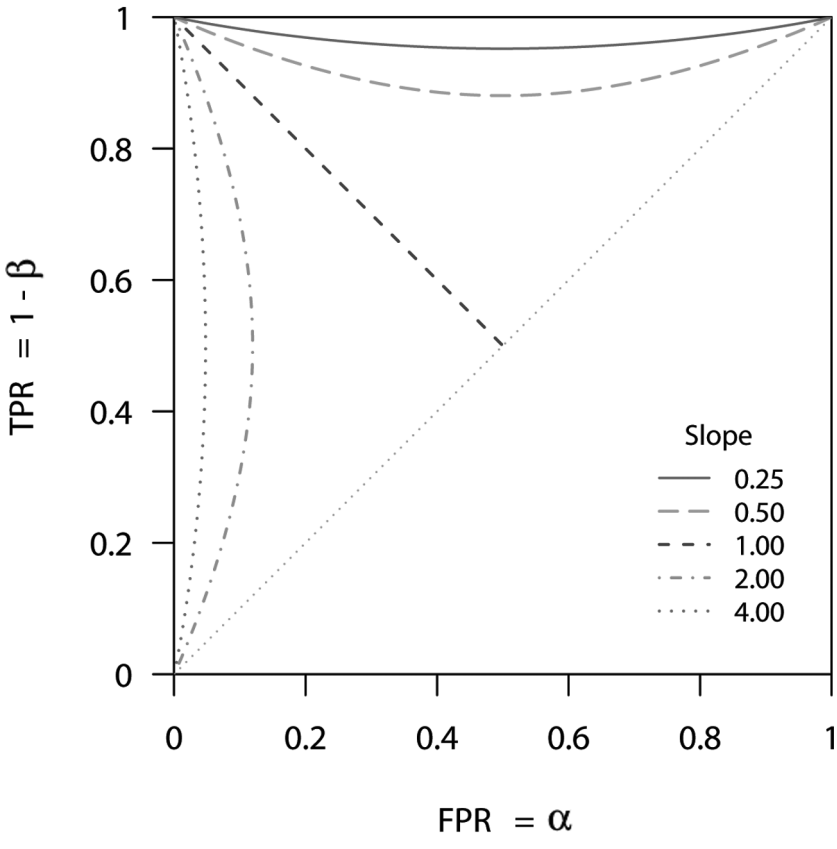


Figure 7. Isoclines traced as AUC varies within the interval (0.5, 1), for different fixed slopes of the iso-expected-cost line. Color version available as an online enhancement.

(i.e., both type I and type II errors). But even above the critical AUC level, value considerations continue to determine the location of the optimal inferential threshold x^* and, thereby, both α^* and β^* as well.

Turning to the second point, the importance of value considerations may be quantified as the relative change in the (β^*/α^*) ratio induced by a unit relative change in the $(p_{\neg D}/P_D)[IC_{\mathrm{FP}}/IC_{\mathrm{FN}}]$ slope. This is a measure of the sensitivity of the (β^*/α^*) ratio to changes in the iso-expected-cost-line (i.e., values) slope. Here we formalize relative changes as percentages, which permit meaningful comparisons, for example, between different AUC levels. Consider two AUC levels AUC′, AUC″ $\in [0.5, 1]$ such that AUC″ > AUC′. The larger the AUC becomes, the more 'bowed' the ROC curve. Consequently, for a 1% increase in $(p_{\neg D}/P_D)[IC_{\mathrm{FP}}/IC_{\mathrm{FN}}]$ the percentage increase

https://doi.org/10.1086/702027 Published online by Cambridge University Press

in (β^*/α^*) is smaller when the resolving power of science is AUC″ rather than AUC′:

$$\left.\frac{\left(\triangle\frac{\beta^*}{\alpha^*}\right)\Big/\left(\frac{\beta^*}{\alpha^*}\right)}{\left(\triangle\frac{p_{\neg D}IC_{\mathrm{FP}}}{p_D IC_{\mathrm{FN}}}\right)\Big/\left(\frac{p_{\neg D}IC_{\mathrm{FP}}}{p_D IC_{\mathrm{FN}}}\right)}\right|_{\mathrm{AUC}''} < \left.\frac{\left(\triangle\frac{\beta^*}{\alpha^*}\right)\Big/\left(\frac{\beta^*}{\alpha^*}\right)}{\left(\triangle\frac{p_{\neg D}IC_{\mathrm{FP}}}{p_D IC_{\mathrm{FN}}}\right)\Big/\left(\frac{p_{\neg D}IC_{\mathrm{FP}}}{p_D IC_{\mathrm{FN}}}\right)}\right|_{\mathrm{AUC}'}. \tag{11}$$

The (β^*/α^*) ratio becomes less sensitive to changes in values $(p_{\neg D}/P_D)$ $[IC_{\mathrm{FP}}/IC_{\mathrm{FN}}]$ as the resolving power of science (i.e., AUC) increases.

7. Conclusion. We have aimed to show that SDT provides a tractable formal structure with which to address questions about where inferential thresholds should be placed, 'reverse engineer' values from choices of inferential thresholds, and clarify conceptual questions regarding inductive risk. Within SDT, optimal inferential thresholds necessarily reflect base-rate information. Yet base rates have received little attention within the AIR literature. Thus, the fact that base rates powerfully modify how much error costs matter for optimal inferential thresholds suggests that this nexus should be addressed within the AIR literature in the future. A near-zero base rate in particular exerts an overwhelming, dominant influence on the location of the optimal inferential threshold.

Furthermore, SDT's formal structure carries implications for a number of ongoing debates concerning inductive risk. In some cases SDT suggests a particular reframing, while in others SDT suggests a particular resolution. As an example of the latter, the SDT-based approach suggests that benefits accruing in each state × inference category need to be accounted for. The consequences of not doing so would include an inability to incorporate opportunity costs, because these are defined as forgone benefits. As an example of the former, the SDT-based approach suggests a way in which we can understand and quantify the notion that values become 'less important' as scientific uncertainty decreases. It turns out that erroneous-inference probabilities (α^*, β^*) generally do not decrease monotonically with the resolving power of scientific experiments, and values continue to influence (α^*, β^*) throughout. However, when framed in terms of sensitivity, indeed the responsiveness of the (β^*/α^*) ratio to changes in values does diminish as the resolving power of scientific experiments increases.

Underpinning these conceptual insights, SDT allows one to solve for the optimal inferential threshold in the underlying score variable x^*, or equivalently in type I (α^*) and type II (β^*) error probabilities. In this standard 'prospective' mode, SDT provides a precise answer to the question of where the inferential threshold should be placed, along with an equally precise answer as to why. It is also straightforward to implement counterfactual analyses, for instance, to determine how much the inferential threshold would shift if the

https://doi.org/10.1086/702027 Published online by Cambridge University Press

resolving power of experiments (d') improved by a specified amount, such as 10% or 20%. Equally, it is straightforward to determine the amount by which a particular parameter, or collection of parameters, would have to change in order for the type I (or alternatively the type II) error probability to fall to a specified target level.

SDT may also be employed retrospectively to 'reverse engineer' the values that rationalize the use of a particular inferential threshold in a specific context. Whereas the inductive-risk literature has discussed the $\alpha = 0.05$ inferential-threshold convention used in many scientific disciplines, with SDT one may reverse engineer the incremental-misclassification-cost ratio required to rationalize the $\alpha = 0.05$ threshold given the base rate and the experiment's discriminability d'. Since both the base rate and discriminability vary greatly from one hypothesis to another, both within a scientific field as well as across scientific fields, rigid application of the $\alpha = 0.05$ inferential threshold entails that most hypotheses are tested using an inferential threshold that is, to a greater or lesser extent, biased. By employing SDT in reverse-engineering mode, this bias may be quantified, mapped, and studied. Moreover, in settings in which inferential procedures are performed on members of the public, SDT in reverse-engineering mode can reveal whose interests are being superseded and whose interests are being accommodated. For instance, does a medical diagnostic threshold reflect the misclassification costs of the patient, the medical practitioner, the clinic or hospital, the health insurance company, or the national health care system? Armed with this information and the SDT formal framework, moral philosophy can address the question of whose interests and misclassification costs should be used in setting diagnostic inferential thresholds.

REFERENCES

Ayles, H., A. Schaap, A. Nota, C. Sismanidis, R. Tembwe, P. De Haas, M. Muyoyeta, N. Beyers, and P. Godfrey-Faussett. 2009. "Prevalence of Tuberculosis, HIV and Respiratory Symptoms in Two Zambian Communities: Implications for Tuberculosis Control in the Era of HIV." *PloS ONE* 4 (5): 1–12.

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis*. Berlin: Springer.

Betz, Gregor. 2013. "In Defence of the Value Free Ideal." *European Journal for Philosophy of Science* 3 (2): 207–20.

Biddle, Justin B., and Rebecca Kukla. 2017. "The Geography of Epistemic Risk." In *Exploring Inductive Risk: Case Studies of Values in Science*, ed. K. Elliott and T Richards, 215–37. Oxford: Oxford University Press.

Bobinac, Ana, Job van Exel, Frans F. H. Rutten, and Werner B. F. Brouwer. 2014. "The Value of a QALY: Individual Willingness to Pay for Health Gains under Risk." *PharmacoEconomics* 32 (1): 75–86.

Cantor, Scott B., Charlotte C. Sun, Guillermo Tortolero-Luna, Rebeccas Richards-Kortum, and Michele Folen. 1999. "A Comparison of C/B Ratios from Studies Using Receiver Operating Characteristic Curve Analysis." *Journal of Clinical Epidemiology* 52 (9): 885–92.

Carbone, Andrea da Silva Santos, et al. 2015. "Active and Latent Tuberculosis in Brazilian Correctional Facilities: A Cross-Sectional Study." *BMC Infectious Diseases* 15 (24): 1–8.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

Churchman, C. W. 1948. "Statistics, Pragmatics, Induction." *Philosophy of Science* 15:249–68.
Cranor, C. 1993. *Regulating Toxic Substances: A Philosophy of Science and the Law*. Oxford: Oxford University Press.
DeBaun, Michael R., and Harold C. Sox. 1991. "Setting the Optimal Erythrocyte Protoporphyrin Screening Decision Threshold for Lead Poisoning: A Decision Analytic Approach." *Pediatrics* 88 (1): 121–31.
DeCarlo, Lawrence T. 1998. "Signal Detection Theory and Generalized Linear Models." *Psychological Methods* 3 (2): 186–205.
Douglas, Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67 (4): 559–79.
———. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.
Egan, James. 1975. *Signal Detection Theory and ROC Analysis*. London: Academic Press.
Elliott, Kevin C. 2011. "Direct and Indirect Roles for Values in Science." *Philosophy of Science* 78 (2): 303–24.
Evans, K. K., R. L. Birdwell, and J. M. Wolfe. 2013. "If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening." *PLoS ONE* 8 (5): 1–6.
Fournet, N., A. Sanchez, V. Massari, L. Penna, S. Natal, E. Biondi, and B. Larouze. 2006. "Development and Evaluation of Tuberculosis Screening Scores in Brazilian Prisons." *Public Health* 20 (10): 976–83.
Green, D. M., and J. A. Swets. 1966. *Signal Detection Theory and Psychophysics*. London: Wiley.
Hagen, Michael D. 1995. "Test Characteristics: How Good Is That Test?" *Primary Care: Clinics in Office Practice* 22 (2): 213–33.
Hempel, Carl G. 1965. "Science and Human Values." In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, 81–96. New York: Free Press.
Jeffrey, Richard C. 1956. "Valuation and Acceptance of Scientific Hypotheses." *Philosophy of Science* 22 (3): 237–46.
Joyce, James M. 1998. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65 (4): 575–603.
Kadane, Joseph B., Mark J. Schervish, and Teddy Seidenfeld. 1999. *Rethinking the Foundations of Statistics*. Cambridge: Cambridge University Press.
Kaivanto, Kim. 2014. "The Effect of Decentralized Behavioral Decision Making on System-Level Risk." *Risk Analysis* 34 (12): 2121–42.
Kik, S. V., S. P. J. Olthof, J. T. N. de Vries, D. Menzies, N. Kincler, J. van Loenhout-Rooyakkers, C. Burdo, and S. Verver. 2009. "Direct and Indirect Costs of Tuberculosis among Immigrant Patients in the Netherlands." *BMC Public Health* 9 (283): 1–9.
Kitcher, Philip. 2001. *Science, Truth, and Democracy*. New York: Oxford University Press.
Levi, Isaac. 1960. "Must the Scientist Make Value Judgments?" *Journal of Philosophy* 57 (11): 345–57.
———. 1962. "On the Seriousness of Mistakes." *Philosophy of Science* 29 (1): 47–65.
Levi, Keith. 1985. "A Signal Detection Framework for the Evaluation of Probabilistic Forecasts." *Organizational Behavior and Human Decision Processes* 36 (2): 143–66.
Lusted, Lee B. 1971. "Decision-Making Studies in Patient Management." *New England Journal of Medicine* 284:416–24.
Marzban, Caran. 2004. "The ROC Curve and the Area under It as Performance Measures." *Weather and Forecasting* 19 (6): 1106–14.
Morrison, Margaret. 2014. "Values and Uncertainty in Simulation Models." *Erkenntnis* 79 (S5): 939–59.
Parker, Wendy S. 2010. "Predicting Weather and Climate: Uncertainty, Ensembles and Probability." *Studies in History and Philosophy of Modern Physics* 41 (3): 263–72.
Pettigrew, Richard. 2016. "Accuracy, Risk, and the Principle of Indifference." *Philosophy and Phenomenological Research* 92 (1): 35–59.
Roberts, C. A., M. N. Lobato, L. B. Bazerman, R. Kling, A. A. Reichard, and T. M. Hammett. 2006. "Tuberculosis Prevention and Control in Large Jails: A Challenge to Tuberculosis Elimination." *American Journal of Preventive Medicine* 30 (2): 125–30.
Rudner, Richard. 1953. "The Scientist *qua* Scientist Makes Value Judgments." *Philosophy of Science* 20 (1): 1–6.

https://doi.org/10.1086/702027 Published online by Cambridge University Press

Scarantino, Andrea. 2010. "Inductive Risk and Justice in Kidney Allocation." *Bioethics* 24 (8): 421–30.
Schervish, Mark J., Teddy Seidenfeld, and Joseph B. Kadane. 2002. "A Rate of Incoherence Applied to Fixed-Level Testing." *Philosophy of Science* 69 (Proceedings): S248–S264.
Shrader-Frechette, K. 1991. *Risk and Rationality: Philosophical Foundations for Populist Reforms*. Berkeley: University of California Press.
Steel, Daniel. 2010. "Epistemic Values and the Argument from Inductive Risk." *Philosophy of Science* 77 (1): 14–34.
———. 2016. "Climate Change and Second-Order Uncertainty: Defending a Generalized, Normative, and Structural Argument from Inductive Risk." *Perspectives on Science* 24 (6): 696–721.
Steele, Katie. 2012. "The Scientist qua Policy Advisor Makes Value Judgments." *Philosophy of Science* 79 (5): 893–904.
Swets, John A. 2001. "Signal Detection Theory, History Of." In *International Encyclopedia of the Social and Behavioral Sciences*, ed. Neil J. Smelser, 14078–82. Amsterdam: Elsevier.
Swets, John A., Robyn M. Dawes, and John Monahan. 2000. "Better Decisions through Science." *Scientific American* 283 (4): 82–87.
Ulehla, Z. Joseph. 1966. "Optimality of Perceptual Decision Criteria." *Journal of Experimental Psychology* 71 (4): 564–69.
Wald, Abraham. 1942. *On the Principles of Statistical Inference*. Notre Dame, IN: University of Notre Dame Press.
Wilholt, Torsten. 2009. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science* 40 (1): 92–101.
———. 2016. "The Seriousness of Mistakes and the Benefits of Getting It Right: Symmetries and Asymmetries in Epistemic Risk Management." Paper presented at the 5th René Descartes Lectures Workshop, Tilburg University, September 5–7.
Winsberg, Eric. 2012. "Values and Uncertainties in the Predictions of Global Climate Models." *Kennedy Institute Ethics Journal* 22 (2): 111–37.
Wolfe, Jeremy M., Todd S. Horowitz, and Naomi M. Kenner. 2005. "Rare Items Often Missed in Visual Searches." *Nature* 435:439–40.
Wolfe, Jeremy M., Todd S. Horowitz, M. J. Van Wert, N. M. Kenner, S. S. Place, and N. Kibbi. 2007. "Low Target Prevalence Is a Stubborn Source of Errors in Visual Search Tasks." *Journal of Experimental Psychology: General* 136 (4): 623–38.

https://doi.org/10.1086/702027 Published online by Cambridge University Press