Normativity, Epistemic Rationality, and Noisy Statistical Evidence

Boris Babic, Anil Gaba, Ilia Tsetlin, and Robert L. Winkler

Forthcoming in *The British Journal for the Philosophy of Science*

March 5, 2021

Abstract

Many philosophers have argued that statistical evidence regarding group characteristics (particularly stereotypical ones) can create normative conflicts between the requirements of epistemic rationality and our moral obligations to each other. In a recent paper, Johnson-King and Babic argue that such conflicts can usually be avoided: what ordinary morality requires, they argue, epistemic rationality permits. In this paper, we show that as data gets large, Johnson-King and Babic's approach becomes less plausible. More constructively, we build on their project and develop a generalized model of reasoning about stereotypes under which one can indeed avoid normative conflicts, even in a big data world, when data contain some noise. In doing so, we also articulate a general approach to rational belief updating for noisy data.

1 Introduction

In a world characterized by socioeconomic and other inequalities, some stereotypes will be statistically sound. In those cases, many philosophers have argued, epistemic rationality can come apart from our moral obligations to each other (e.g., Basu and Schroeder, [2019]). For example, Tamar Gendler puts the point as follows:

> As long as there's a differential crime rate between racial groups, a perfectly rational decision maker will manifest different behaviors, explicit and implicit, toward members of different races. This is a profound cost: *living in a society structured by race appears to make it impossible to be both rational and equitable* (Gendler, [2011], p. 57, emphasis added).

This is an example of what has come to be known as normative conflict (Basu, [2018]). It refers to the notion that in an unequal society we often learn about the uneven distribution of stigmatized traits along sensitive demographic lines. We are then, it seems, forced by the hand of epistemic rationality to formulate beliefs about vulnerable groups that strike many people as immoral.

In a recent paper, Johnson-King and Babic ([2020]) (JKB henceforth) argue that such normative conflicts can ordinarily be avoided: what ordinary morality demands, they argue, epistemic rationality typically permits. They rely on the notion of minimizing epistemic risk, developed in Babic ([2019]), as a principle for identifying an appropriate prior which captures the relevant normative considerations at stake. As a result, they avoid normative conflicts by explaining how in most such cases epistemic rationality permits a much wider set of priors than has previously been assumed.

In this project, however, we explain that as a dataset gets large (in the sense of the number of observations in a sample), relying on moral attitudes to restructure the prior becomes an increasingly less sensible way to avoid normative conflicts. With very large samples, this strategy requires very stubborn priors which ultimately undermine an agent's ability to learn. More constructively, we build on their project and develop a generalized model of reasoning about stereotypes under which one can indeed avoid normative conflicts, even in a big data world, when such data contain some noise. In doing so, we also articulate a model of rational belief updating in response to learning experiences characterized by large but noisy samples.

The paper proceeds as follows. First, we explain the basic notion of epistemic risk and briefly describe the argument in JKB. A key step in their argument is that different attitudes to epistemic risk license different priors in the absence of other information. And in most cases giving rise to normative conflicts, there will exist an epistemically per-

missible prior which cautions an agent from adopting stereotype reinforcing credences (for instance: a prior which cautions against adopting a high credence that members of some racial groups are more likely to commit certain crimes, in Gendler's example). Second, we articulate our challenge to this argument: with large datasets, tweaking the priors only goes so far – the likelihood dominates inferences and normative conflicts reemerge. Third, and most importantly, we develop a model of belief updating under noise and use it to explain how such normative conflicts can be avoided still, when the data is not perfect. Our model leaves room for identifying true population differences where they exist.

2 Priors and Epistemic Risk

We develop the argument to follow within the general framework of epistemic utility theory (see e.g., Joyce, [1998]; Pettigrew, [2016]). In particular, we assume that an epistemically rational agent should adopt credences in a way that minimizes expected inaccuracy, where inaccuracy is measured by an appropriate scoring rule. Generally, a scoring rule is appropriate if it is monotonic, continuous, and strictly proper.[1] These properties, together with some modest decision theoretic norms, commit us to probabilism – the thesis that subjective credences should conform to the probability axioms – and conditionalization – the thesis that upon receiving new information (in ordinary circumstances),[2] one should update by Bayes' Rule (Joyce, [2009]; Greaves and Wallace, [2006]). This, in a nutshell, is what we take to be a floor on epistemic rationality.

Consider a dichotomous proposition, 'A'. For example, 'Alice will be tenured next year'. The core idea in Babic ([2019]) is that there are two ways of being inaccurate about A. We may increase our credence in A, $p(A)$, when A is false. This is a case of increasing inaccuracy in the false positive direction, so to speak. Or we may decrease our credence in A, $p(A)$, when A is true. This is a case of increasing inaccuracy in the false negative direction. Let $s(p(A), I_A)$ be a measure of the inaccuracy of the credence for A, where I_A is an indicator variable that equals 1 if A is true and 0 otherwise (note that higher accuracy is equivalent to lower inaccuracy). For scoring rules that are symmetric, in the sense that $s(p, 1) = s(1 - p, 0)$ for all $p \in [0, 1]$, a unit increase in inaccuracy in either direction should be treated equally. But many continuous, monotonic, strictly proper scoring rules are not symmetric and, indeed, we may well care differently about these two error directions regarding A.

[1] For discussion of these properties, we refer the reader to Joyce ([2009]) and Babic ([2019]).

[2] There are of course disputed cases, such as when evidence is uncertain, conditional, or non-propositional, but no such cases will arise here. See, for example, van Fraassen ([1981]) and Howson and Franklin ([1994]).

For example, if we are concerned about offending Alice by assuming she is less talented than she is, then false confidence in her denial of tenure is relatively worse. Alternatively, if we are especially concerned about helping Alice decide whether to go on the job market while awaiting tenure, then falsely assuring her she will get tenure might be relatively worse. Figure (1) depicts a pair of scoring rules to illustrate the point. The scoring rule on the left is symmetric, and penalizes increases in inaccuracy (y-axis) in either direction equally, whereas the scoring rule on the right punishes increases in inaccuracy in the false negative direction less. The way one adjudicates the relative costs of increasing inaccuracy in either direction will determine the scoring rule they deem appropriate.

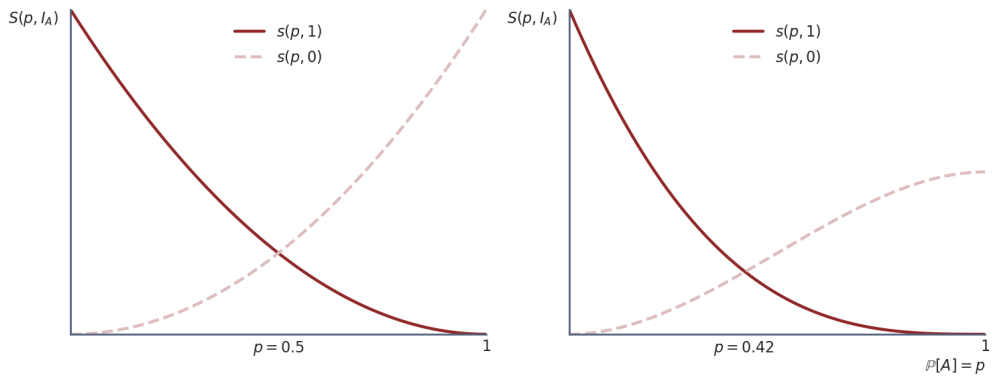

Figure 1: Symmetric scoring rule (left), and asymmetric scoring rule (right).

Because the scoring rules are continuous, and decreasing (increasing) in $p(A)$ when A is true (false), the intermediate value theorem guarantees that the point of intersection we see in both plots must exist for reasonable measures of inaccuracy. This is the point at which there is no variability in inaccuracy outcomes – the safest point, or the point of zero epistemic risk. Call it p^*. More generally, Babic ([2019]) defines the epistemic risk associated with investing probability p in A as

$$R(p) = \int_{p}^{p^*} |s(t, 1) - s(t, 0)| dt \tag{1}$$

when $p < p^*$. When $p > p^*$ the bounds of integration are reversed. Thus, if we want to adopt a maximally safe prior for A, in terms of inaccuracy, we should adopt $p(A)$ which satisfies $s(p, 1) = s(p, 0)$. In this way, we can use the notion of epistemic risk, and the associated normative attitudes it incorporates, in order to identify reasonable

priors in the absence of information. In the example above, an epistemic risk minimizer who cares about errors symmetrically would adopt $p = 0.5$, whereas an epistemic risk minimizer whose attitudes to error correspond to the scoring rule on the right would adopt $p = 0.42$.

This recipe provides a family of indifference principles, so to speak, for identifying priors – with each instance corresponding to different temperaments about the relative severity of increasing inaccuracy. In the next section, we will explain how JKB rely on this idea in order to avoid normative conflicts.

3 Normative Conflicts

Johnson-King and Babic ([2020]) apply the notion of epistemic risk in order to explain how one can avoid normative conflicts. They use the following example.

> **Gender Bias Study**. One morning, you read a report about a study on gender discrepancies in academic employment. The study surveyed 500 men and 500 women employed in universities. They found that only 30% of the women were employed in faculty positions, while the other 70% were administrative assistants. For men, the proportions were reversed. Before learning this, you had no prior relevant information. The study was otherwise legitimate. You then meet Mary. Mary tells you she works in a university. What should be your credence that Mary is a faculty member?

Many epistemologists would argue that epistemic rationality requires one to believe it is 70% likely that Mary is an administrative assistant. This conclusion would follow from the so-called frequency-credence connection, and lead directly to normative conflict. Call this the naive answer.

JKB argue against the naive answer by using the standard Bayesian approach to predictive inference, but structuring the prior in their model by taking into account attitudes to epistemic risk. We will develop their model with care here, as we build on it in the next section.

Let θ be the (unknown) proportion of women in academia who are faculty. Suppose that in a sample of n women in academia, we observe x women who are faculty. Then x follows a Bernoulli process with parameter θ and the likelihood function is given by

$$\ell(x|\theta, n) = \theta^x(1-\theta)^{n-x} \tag{2}$$

In the Bayesian approach, we need to identify prior beliefs regarding θ. A beta distribution, being relatively flexible, can approximate a wide variety of information states

regarding a Bernoulli process and is commonly used in Bayesian models involving proportions (Lindley and Phillips, [1976]). Let $f(\theta)$ be the prior probability density for θ, where

$$f(\theta) = f_\beta(\theta|a_\theta, b_\theta) = \theta^{a_\theta - 1}(1-\theta)^{b_\theta - 1}/B(a_\theta, b_\theta) \tag{3}$$

is a beta density function with $B(a_\theta, b_\theta) = \Gamma(a_\theta)\Gamma(b_\theta)/\Gamma(a_\theta + b_\theta)$. The mean and variance of a beta distribution are given by $\mathrm{E}[\theta] = a_\theta/(a_\theta + b_\theta)$ and $\mathrm{Var}(\theta) = a_\theta b_\theta/[(a_\theta + b_\theta)^2(a_\theta + b_\theta + 1)]$. As a_θ becomes larger the distribution moves towards the right, whereas an increase in b_θ moves the distribution towards the left. When $a_\theta = b_\theta$, the distribution is symmetric around 0.5. If both a_θ and b_θ increase the distribution begins to narrow. The parameters a_θ and b_θ can also be interpreted as "pseudo" observations upon which the prior beliefs are based. For instance, $a_\theta = 7$ and $b_\theta = 3$ would be equivalent to having observed 7 faculty and 3 non-faculty out of 10 women in academia. These properties will be important to our argument in Section 4. Note, also, that as $(a_\theta + b_\theta)$ increases, the prior becomes more resilient, in Joyce ([2005])'s sense – it exerts more weight on the posterior.

With the likelihood in (2) and the prior in (3), the posterior density for θ is given by

$$f(\theta|x, n) = f_\beta(\theta|a_\theta + x, b_\theta + n - x) \propto \theta^{a_\theta + x - 1}(1-\theta)^{b_\theta + (n-x) - 1} \tag{4}$$

Note that the posterior distribution is of the same form as the prior distribution. This is because a beta distribution is *conjugate* to the Bernoulli process. This means that if we start with a beta prior for θ, and update via Bayes' Rule with data from a Bernoulli process, our posterior will likewise be beta but with updated parameters. Such a model lends itself to an intuitive interpretation. The posterior beta distribution for θ after seeing the data is given by adding the actual observations (in the sample data) and the corresponding pseudo observations represented in the prior distribution. For example, suppose our prior for θ is a beta density with parameters $(a_\theta = 7, b_\theta = 3)$, and we observe 4 out of 10 women in academia who are faculty. Our posterior for θ would be a beta density with parameters (7+4, 3+6).

But in order to formulate a credence about Mary, we need more than the posterior distribution. Let $\widetilde{X} \in \{0, 1\}$ be an additional outcome that has yet to be observed (i.e., Mary). The distribution of $\widetilde{X}$ given x is called the predictive distribution, and is of the following form:

$$P(\widetilde{X} = 1|x) = \int_0^1 P(\widetilde{X} = 1|\theta, x) f(\theta|x) d\theta \;\; = \frac{a_\theta + x}{a_\theta + b_\theta + n}. \tag{5}$$

Huttegger ([2017]) refers to the expression in (5) as the generalized rule of succession and shows that this form of the predictive probability follows from several modest

assumptions about the structure of the data-generating process, which are satisfied here. The question for us is: which values should we assign to a_θ and b_θ?[3]

First, the naive answer requires that $a_\theta = b_\theta = 0$. This would result in an improper (and arbitrarily incoherent) prior, because $1/(\theta(1-\theta))$ is not bounded. Second, following Laplace's rule of succession, we could set $a_\theta = b_\theta = 1$. This is equivalent to a uniform prior for θ. And third, JKB follow Huttegger's generalized rule of succession, but they use attitudes to epistemic risk in order to identify appropriate values for a_θ and b_θ – i.e., to make specific the generalized rule of succession. Suppose we find it more costly to falsely increase confidence in Mary not being a professor (false negative mistake) than we do to falsely increase confidence in Mary being a professor (false positive mistake). Then we need a a scoring rule which penalizes inaccuracy more in the direction of false negative mistakes, and the safest (risk-free) point will be above 0.5. If we wish to minimize epistemic risk, then we should choose a beta prior which is such that $\mathrm{E}[\theta] = p^*$ where p^*, as noted above, satisfies $s(p,1) = s(p,0)$. It is now a short step to see how JKB seek to avoid normative conflicts. In particular, we can state an explicit condition on what kind of prior might be "morally required," so to speak, as follows.

> **JKB Approach.** If one wants the posterior probability (that, say, Mary is a professor) to be above 0.5 due to underlying normative considerations then, if the sample mean x/n is less than 0.5, it must be the case that $\alpha_\theta - \beta_\theta > n - 2x$.

Under this approach, one's prior, selected by consulting the underlying normative considerations, will guarantee that the updated credence for the claim that, say, Mary is faculty, will be above 0.5. JKB's argument is that there is no obvious reason that this prior is epistemically impermissible – the agent is coherent and is updating by Bayes' Rule. If one finds the agent's prior objectionable, an argument is needed that such distorted attitudes to error (or, rather, to epistemic risk) are unreasonable in the particular context. One cannot appeal, without circularity, to something like the ordinary principle of indifference, because that principle tacitly presupposes attitudes to epistemic risk – namely, that both error directions are equally costly.

[3]Note that $(a_\theta + x)/(a_\theta + b_\theta + n)$ approaches the sample mean, x/n, as x and n go to infinity. This implies that the prior washes out in the limit. But this is also why, on the JKB approach, as $n \to \infty$, $(\alpha_\theta + \beta_\theta)$ also has to go to infinity. We explain this further below. Thanks to an anonymous reviewer for emphasizing this point.

4 Stereotypes Under Noisy Big Data

The JKB Approach explains many ordinary cases of normative conflict. Like the Gender Bias Study, if all one has to go on is a newspaper description of an isolated report, perhaps it is worth being extra careful in applying it to Mary. And indeed, the reason we are hesitant to jump onto stereotype encoding evidence is precisely because we are worried about the cost of being wrong if the individual in question proves to be an exception.

But notice the strategy employed here: we avoid a pernicious prediction by setting up our prior in a way that hedges against it, and we justify this due to the asymmetry of the cost of mistakes in the relevant problem. While this can be an appropriate way to reflect differential costs of error, with increasingly larger samples we become vulnerable to the objection that we are burying our head in the sand, so to speak. We have to load our prior to such an extent that we become increasingly stubborn and unable to learn. And the implied normative attitudes to error that are required to carry this heavy burden begin to look epistemically unreasonable. If instead of observing 500 women, we observe 100,000 women, the JKB approach would require a prior that is almost arbitrarily peaked around 0.7 in order to avoid normative conflict. The figure below illustrates this point for a sample size of $5,000$, $10,000$, and $100,000$. In each case, the sample mean (of female faculty) is assumed to remain as in the original hypothetical, i.e., 0.3.

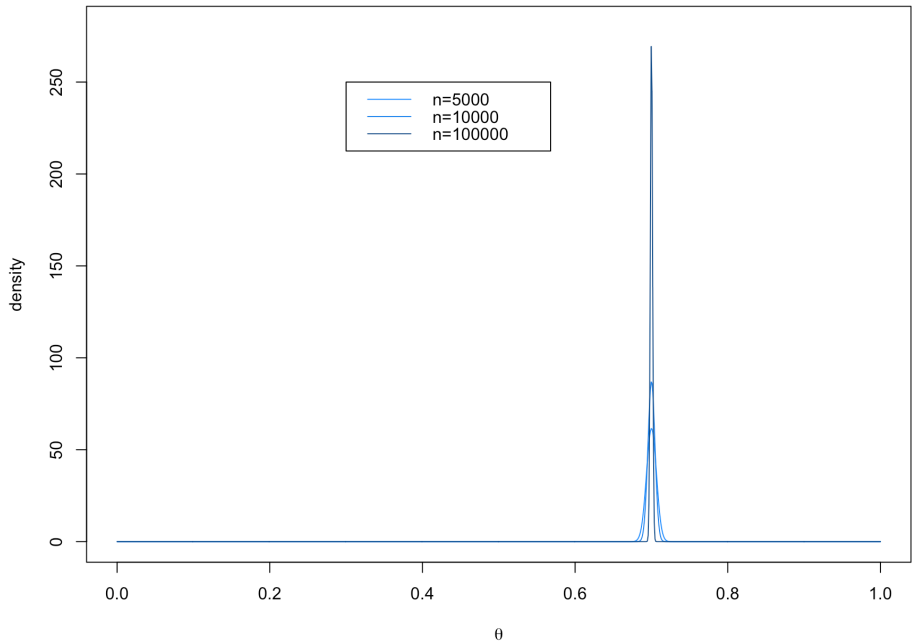

Figure 2: Prior distributions for θ, if one is to avoid normative conflict, for $n = 5000$, $n = 10000$, $n = 100000$.

The reason that such sharply peaked priors undermine one's ability to learn is, as mentioned, because if one seeks a posterior probability (that, say, Mary is a faculty member) to be above 0.5 due to underlying normative considerations then, if the sample mean x/n is less than 0.5, $(\alpha - \beta)$ must be greater than $n - 2x$. The pseudo count increases linearly in n, which means that for very large sample sizes, the prior that is called for by normative considerations will become extremely dogmatic. Or, as Joyce ([2005]) calls it, resilient – but resilient to a fault. For example, if we start with a Beta$(1000, 1000)$ prior for a coin's bias and observe 19 heads and 1 tails, our posterior point estimate would be 0.504. Indeed, even if we toss the coin 100 times and observe *all* heads, our posterior point estimate would be 0.52 even though the probability of observing 100 heads conditional on a fair coin is 0.5^{100}. Intuitively, 100 tosses of a coin with every single one of them heads is extremely strong evidence against the hypothesis that the coin is fair, yet this agent barely moves away from the estimate that it is exactly fair.

While the reemergence of normative conflicts is perhaps inevitable with perfect data, large datasets are rarely perfect. When we collect statistical evidence about people, such as in the Gender Bias Study, there typically exists some noise in the data.

And, the level of noise is usually unknown. This can lead to unobservable misclassification error in the data. Suppose, for example, we survey people about their voting intentions, asking whether they will vote Democrat or Republican in the next presidential election. The observed data in the survey may differ from actual voting behavior for a variety of reasons. Some respondents in the survey may intentionally misreport, may be leaning toward voting for one candidate but still somewhat undecided, or may change their mind at the time of voting in light of new information. These are all in addition to the possibility of the responses being coded incorrectly. The figure below gives a couple of examples of how the actual voting behavior might differ from reported voting intentions.

	Republican	Democrat
Reported	50	50
Actual (one possibility)	55	45
Actual (another possibility)	45	55

Table 1: Sampling error in voting.

This is not surprising. We are all too familiar with such voting analyses. More generally, however, noise of this sort is the norm rather than the exception when we collect data. Insofar as the Gender Bias Study assumes perfect sampling, therefore, it is not representative of the types of cases we usually face in ordinary decision making. Accordingly, consider a simple case which mirrors the Gender Bias Study but where the sample is large and the possibility of error exists.

> **Recruitment**. A study on gender disparities in performance among investment bankers looked at evaluations of 100,000 men and 100,000 women in junior positions. The evaluations recorded each employee as "director worthy" or "not director worthy" (talented and untalented, for short). The researchers found that only 30% of the women were recorded as talented whereas for men, 70% were recorded as talented. The research is otherwise sound. Later, you meet Alice, who has applied for a job at your bank. How confident are you that Alice is **actually** talented, based on data regarding the proportion of women who were recorded, or **deemed**, talented?

Using the JKB recipe, the probability would be proportional to $p(\theta)f(\mathbf{x}|\theta)$ where θ follows a beta distribution, with a_θ and b_θ being determined by the agent's attitudes to epistemic risk, and each x is a Bernoulli draw (corresponding to each woman being talented or untalented). To avoid normative conflicts, the prior would have to be arbitrarily peaked, as Figure (2) illustrates. Accordingly, the JKB model is not appropriate

for a case like Recruitment, because it ignores the possibility that an actually talented woman is recorded as untalented.

We should emphasize that when we say "recorded" as untalented, we are not referring merely to clerical errors. Rather, we mean to denote the situation where the true but unobservable state (person is talented) comes apart from the state imputed to them (person is deemed to be not talented). Consider a simple example. The LSAT score measures one's aptitude for success in law school. But aptitude is never observed, and LSAT performance is only an imperfect indicator of it. It is possible for someone who would be a very successful law student to score poorly on the LSAT. This could of course happen because their actual score is incorrectly recorded. But it could also happen that their score does not correspond to their aptitude, due to contingent testing circumstances (anxiety, illness, lack of sleep, etc.), or due to more general underlying socioeconomic inequities that make them less well prepared as compared to their peers to sit for the LSAT without affecting their aptitude for legal study. We will restructure the likelihood in the JKB model so as to create logical space for this kind of discrepancy between actual aptitude (i.e., potential to be a successful law student) and observed aptitude (i.e., LSAT score).

Turning back to the Recruitment example, this kind of misclassification is a substantial risk in a world we know to be characterized by gender inequality, in which women face especially large barriers to success, and in which implicit bias regarding gender affects women's performance assessments. We can of course disagree on the extent of this, but we do not want to assume its absence by hypothesis. Thus, we need an altogether different model that makes room for the possibility of misclassification. And as we will see below, when we build a model that accommodates this possibility, we find that it is very rarely the case that epistemic rationality calls for stereotype enforcing beliefs. The following material will be somewhat technical, so before we proceed let us explain the motivation, which is quite intuitive.

Recall that in the beta-binomial model, updating beliefs by Bayes' Rule is equivalent to counting the number of favorable and unfavorable observations. Thus, when there is no noise, as in the Gender Bias Study, to compute the posterior we simply add the number of women faculty to our initial value of a_θ, and the number of women non-faculty to our initial value of b_θ. And to make predictions about Mary in the Gender Bias Study example, we simply use the mean derived from these values. When there is noise, as in Recruitment, we don't want to do this, because we suspect the number of women deemed untalented is inflated. Therefore, given some noise rate, and a sample size, what we want to figure out is the "noise-adjusted" sample size, and update on *that* instead before making predictions about Alice. What will be interesting to observe is

how rapidly the noise-adjusted sample size decreases even with modest possible levels of noise and as the uncertainty about the level of noise increases.

As before, let $\theta \in [0, 1]$ be the proportion of women who would be truly successful under ideal conditions, i.e., conditions identical to those of men (or as an alternative interpretation, those who would be promoted absent any socioeconomic, occupational, political etc. gender disparities among men and women). The point is, θ represents women's true talent rate – it is an unobserved latent variable, much like IQ, EQ, or any other indicator of aptitude in education or the workforce. Further, let $\lambda \in [0, 1]$ be the misclassification rate of talented women as untalented. We assume for now that the other type of misclassification (untalented women recorded as talented) is so small that it is not worth worrying about (In Appendix B we expand the model to accommodate both types of error). Then the data generating model we have looks like this:

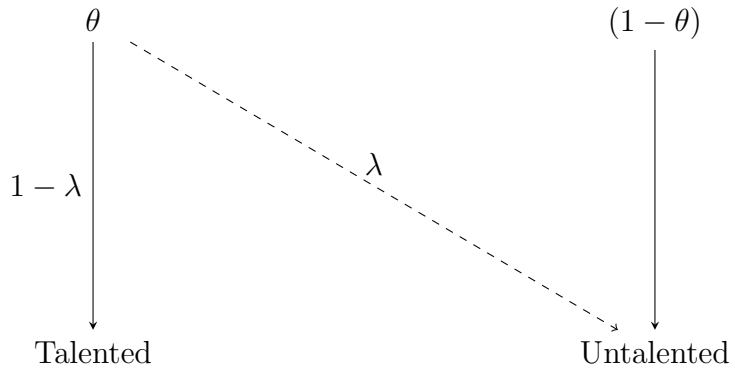

Figure 3: Recruitment with noise level λ.

In this model, the probability that a woman is deemed talented is $\phi = \theta(1 - \lambda)$. The probability that a woman is deemed untalented is $1-\phi = (1-\theta)+\theta\lambda$. In a sample of n women, let γ be the number of women who are deemed talented and the remaining $n - \gamma$ as untalented. Then the data generating process for the recording (as opposed to the actual number) of women as talented and untalented is Bernoulli in ϕ and not in θ (the actual proportion of talented women, as in the JKB model). The likelihood of the sample is thus of the form

$$\ell(\gamma|n, \theta, \lambda) = \big[\theta(1 - \lambda)\big]^{\gamma}\big[(1 - \theta) + \theta\lambda\big]^{n-\gamma}. \tag{6}$$

The maximum likelihood estimate of (θ, λ) is not unique. The likelihood function is unable to distinguish among all (θ, λ) pairs with the same value of ϕ. As a result of this identification problem, the maximum likelihood estimate of (θ, λ) consists of

all (θ, λ) pairs such that $\theta(1-\lambda) = \gamma/n$. For example, if 30% of 100,000 women are observed as talented, then the likelihood is maximized at infinite combinations of (θ, λ) including, for example, $(\theta = 0.3, \lambda = 0), (\theta = 0.5, \lambda = 0.4), (\theta = 0.8, \lambda = 0.625)$ and $(\theta = 1, \lambda = 0.7)$. In other words, many disparate explanations of the data seem equally compelling. If we assume that $\lambda = 0$ then this model is the same as the JKB noise-free model.

The likelihood function in (6) can be expressed as

$$\ell(\gamma|n, \theta, \lambda) = \sum_{t=0}^{n-\gamma} \binom{n-\gamma}{t} \theta^{n-t}(1-\theta)^t \lambda^{n-\gamma-t}(1-\lambda)^\gamma. \tag{7}$$

Here t can be interpreted as the number of women who are correctly classified as untalented. Since we do not know or observe t, the likelihood is expressed as a mixture of the $n-\gamma+1$ likelihoods that could arise with each possible number of misclassifications in the data.

While λ is unknown, one might have some prior beliefs on λ along with those on θ. We might use our background knowledge about social inequality, gender stereotypes in finance, barriers to success for women in business, historical practices of discrimination, and so forth. For instance, *a priori*, $(\theta = 0.7, \lambda = 0.3)$ might be considered more likely than $(\theta = 0.49, \lambda = 0)$ or $(\theta = 0.98, \lambda = 0.5)$, although all three pairs have identical likelihoods for any given data. The Bayesian approach encourages us to start with whatever prior information we can muster. Such beliefs can be expressed in the form of a joint distribution for θ and λ, and given a sample, they can be updated by Bayes' Rule.

Bayesian models with unknown misclassification rates in dichotomous data have been developed in Winkler and Gaba ([1990]), Gaba and Winkler ([1992]), and Gaba ([1993]). For ease of exposition, we restrict attention to their special case with a prior density on (θ, λ) which assumes θ and λ are independent and is given by

$$\begin{aligned} f(\theta, \lambda) &= f_\beta(\theta|a_\theta, b_\theta) f_\beta(\lambda|a_\lambda, b_\lambda) \\ &\propto \theta^{a_\theta - 1}(1-\theta)^{b_\theta - 1} \lambda^{a_\lambda - 1}(1-\lambda)^{b_\lambda - 1}, \end{aligned} \tag{8}$$

where f_β is a beta density as defined in Eq. (3). Restricting attention to the independent case is reasonable in our setting, and not merely a simplifying assumption. One can think of other contexts where such independence might be less valid. Suppose, for example, we are collecting information about self-reported marijuana use among the public. Suppose we do this survey in a city where marijuana use is widespread and attitudes to it are quite liberal. In this city, the misclassification rate is likely to be

fairly low because our respondents will not be worried about admitting to their use habits. Now suppose we take the same survey in a very conservative city where the use of marijuana is very taboo and its use is limited to small, marginalized communities. In this case, the misreporting rate will almost certainly be higher. In general, as a region's attitudes to the use of marijuana liberalize, we can expect that people's dishonesty about their use of it will decrease. So in a case like this, λ is very closely connected to θ, and they move together. But in our case, the prejudice, implicit biases, historical patterns of discrimination and systemic institutional barriers that deflate women's perceived aptitude are not as strongly connected to their true talent rate. In other words, the problem of gender imbalance cannot be solved by simply telling women to do better or work harder.

Admittedly, it is unlikely that the actual correlation between θ and λ would be exactly zero.[4] If women are frequently falsely misclassified as untalented, this may have adverse effects on their confidence and self-perception in a way that to some extent creates a vicious self fulfilling prophecy and affects their actual performance. Likewise, if women are rarely falsely misclassified as untalented, this can have a positive reinforcing effect. There are certainly many ways to model the situation where the observed rate is connected to noise, if this is desired, and when we model such dependence in the data-generating model, it can yield exchangeable data or not. Gaba ([1993]) develop one such model. For example, we can build a hierarchical model, with a separate θ_i for each person and a higher-order distribution for θ from which the θ_i values are drawn. After seeing new data, we would then revise both the higher-order distribution and the distribution for each θ_i. Alternatively, one can consider using different likelihood functions for combining θ and λ. For instance, our model assumes that the probability of being classified as talented is $\theta(1-\lambda)$, in which case the probability decreases proportional to θ. But this is not the only way for λ to disturb θ. For example, we could have $\theta^{1/(1-\lambda)}$. Here, if $\lambda = 0$, there is no disturbance, and if θ is close to 1, the disturbance is small; while for small θ it is very large.

Either alternative could be an interesting extension for future work, but for now we believe we can generate sufficiently rich insights with a relatively parsimonious yet still realistic model. And after all, even if some proportion of the discrepancy turns out to be due to the fact that the minority group is under performing, our point would be that a substantial amount of the observed discrepancy might not be due to actual underlying differences in aptitude but rather to the social conditions that affect our perceptions of it. In other words, the key is that rational inferences about performance across groups should reflect prior uncertainty about λ rather than assuming a priori

[4]We are indebted to two anonymous reviewers for raising the insightful suggestions described in this paragraph, including the alternative likelihood mentioned below.

that $\lambda = 0$.

With the prior in (8), and the likelihood in (7), the posterior density is given by

$$\begin{aligned} f(\theta, \lambda|\gamma, n) &= \sum_{t=0}^{n-\gamma} f(t, \theta, \lambda|\gamma, n) \\ &= \sum_{t=0}^{n-\gamma} f(t|\gamma, n) f(\theta, \lambda|\gamma, n, t) \\ &= \sum_{t=0}^{n-\gamma} w_t f(\theta, \lambda|\gamma, n, t), \end{aligned}$$

where

$$\begin{aligned} w_t &= a_t / \sum_{t=0}^{n-\gamma} a_t, \\ a_t &= \binom{n-\gamma}{t} \mathrm{B}(a_\theta^*, b_\theta^*) \mathrm{B}(a_\lambda^*, b_\lambda^*), \\ f(\theta, \lambda|\gamma, n, t) &= f_\beta(\theta|a_\theta^*, b_\theta^*) f_\beta(\lambda|a_\lambda^*, b_\lambda^*), \end{aligned} \tag{9}$$

with

$$\begin{aligned} a_\theta^* &= a_\theta + n - t, b_\theta^* = b_\theta + t, \\ a_\lambda^* &= a_\lambda + n - \gamma - t, \text{and } b_\lambda^* = b_\lambda + \gamma. \end{aligned}$$

The posterior density in (9) is a mixture of densities of the same form as in (8). The weight w_t is the posterior probability that t out of $n - \gamma$ women recorded as untalented were correctly recorded. And the posterior density is a mixture of $n - \gamma + 1$ possible posterior densities that would result under perfect knowledge of the exact number of misclassifications (i.e., under perfect knowledge of $n - \gamma - t$). Indeed, this expression provides an intuitive explanation for the phenomenon we will soon observe – which is the rapidly diminishing information value of data with even slight additions of noise. This occurs because the posterior is a weighted mixture of many posteriors, one for each possible misclassification. The marginal posterior densities for θ and λ can be obtained as

$$\begin{aligned} f(\theta|\gamma, n) &= \sum_{t=0}^{n-\gamma} w_t f_\beta(\theta|a_\theta^*, b_\theta^*) \\ \text{and} \\ f(\lambda|\gamma, n) &= \sum_{t=0}^{n-\gamma} w_t f_\beta(\lambda|a_\lambda^*, b_\lambda^*). \end{aligned} \tag{10}$$

Now we can return to Recruitment. Recall that we are uncertain about the actual proportion of talented women, but perhaps we start with the assumption that $\theta \sim f_\beta(\theta|7,3)$. This implies that a priori $\mathrm{E}(\theta) = 0.7$, but the distribution is fairly spread out and admits all values of θ between 0 and 1. At the same time, we suspect that there is a meaningful chance of women being incorrectly recorded as untalented in any data that we might see. Suppose that our uncertainty about the one-sided misclassification rate (false negative) is best represented by $\lambda \sim f_\beta(\lambda|3,7)$. Then, $\mathrm{E}(\lambda) = 0.3$, and as in the case of θ, the prior distribution for λ is also quite spread out, admitting values close to zero and as high as 0.7. For both parameters, the sum $(\alpha + \beta)$ is modest, and thus the prior is not objectionably stubborn. We are lightly structuring each prior, which could reflect either prior information, or our attitudes to epistemic risk, or both.

Now suppose, as in the Recruitment example, we observe data on 100,000 women of whom 30% are recorded as talented. Given our model, one thing is immediately clear: the number of women who are actually untalented is inflated in the recorded data. Our built-in conjecture is that this is the result of historical patterns of discrimination and implicit biases leading to women being perceived as on average less talented. But, we are uncertain as to what extent.

It is worth emphasizing that we do not intend to make any essentialist claim between gender and performance here. Indeed, we could equivalently express differential outcomes using either gender, or one of its many correlate characteristics. As in Simpson's Paradox, if we introduce other factors, like years of work experience or education level, we may see a reduction in the differential.[5]

In any case, using our model, we can find the posterior distributions for θ and λ. As can be seen from the above expressions, the analytical calculations require computing combinatorial terms with large values. Hence, we use stochastic simulation to approximate the posterior marginal densities. To draw simulations, our data-generating model for Recruitment can be summarized as follows:

[5]Thanks to an anonymous reviewer for highlighting this point.

$$
\begin{aligned}
\lambda &\sim \text{Beta}(3, 7) \\
\theta &\sim \text{Beta}(7, 3) \\
\phi &= \theta(1 - \lambda) \\
\gamma &\sim \text{Binomial}(\phi, n) \\
n &= 100,000 \\
\frac{1}{n} \sum_{i=1}^{n} \gamma_i &= 0.3
\end{aligned}
$$

Now that the model is specified, we can estimate the joint posterior distribution for (λ, θ) and derive the marginal distributions for each. Figure 4 illustrates the marginal posterior densities for θ and λ that we obtained given the data ($\gamma = 30,000$ and $n = 100,000$), using a Markov Chain Monte Carlo algorithm known as the Gibbs Sampler (Plummer, [2003]; Geman and Geman, [1984]).

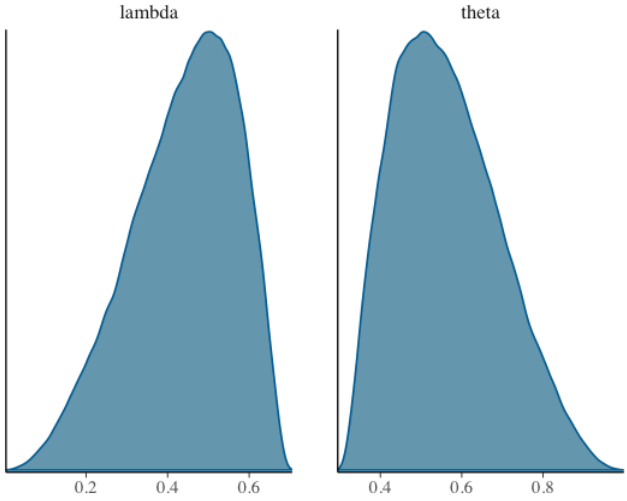

Figure 4: Posterior distributions for θ and λ for $n = 100000$.

Recall that the marginal posterior mean for θ is our predictive probability that Alice in our Recruitment example is talented, using Huttegger's generalized rule of succession. The posterior mean for θ that we obtained using this simulation is 0.55. As a result, despite the very large noisy sample suggesting that only 30% of the women

are talented, our prediction that Alice is talented is very nearly 0.5. In other words, we avoid normative conflict.

To get further insight concerning this model, consider the entire posterior density for θ which provides the full representation of uncertainty about θ, and compare it to a noise-free model. Ignoring misclassifications, i.e., assuming $\lambda = 0$, the prior $f(\theta) = f_\beta(\theta|a_\theta = 7, b_\theta = 3)$ would be revised to the posterior $f(\theta|\gamma, n) = f_\beta(\theta|a_\theta + \gamma = 30007, b_\theta + n - \gamma = 70003)$. Note that the prior mean of θ is 0.7. The noise-free posterior density has a mean of 0.3, with a standard deviation of 0.001, placing almost the entire probability mass at $\theta = 0.3$ and completely overwhelming the prior on θ. On the other hand, in our model with noise, the posterior mean of θ is near 0.5, with a posterior standard deviation of 0.114 which is 114 times larger than in the noise-free case. This is because consideration of all the possible misclassifications that could have occurred in the sample leads to much greater uncertainty about θ. In fact, we can calculate the noise-free sample size leading to the same posterior mean and standard deviation as a noisy sample size. This helps us to see just how much and how quickly the value of information diminishes with noise.

> **Effective sample size**. Given a prior, and a sample of size n, with misclassification rate λ, the effective sample size n^* is the sample which would have the same effect on the prior if $\lambda = 0$.

We compute effective sample size through a matching of moments approach – in particular, by matching posterior variance with n observations and noise in λ to posterior variance with n^* observations and zero noise, as follows.

$$\begin{aligned} \text{Var}(\theta) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{\left(\frac{\alpha}{\alpha+\beta}\right)\left(\frac{\beta}{\alpha+\beta}\right)}{\alpha+\beta+1} \\ &= \frac{\text{E}[\theta][1-\text{E}[\theta]]}{n+1} \\ \rightarrow n^* &= \frac{\text{E}[\theta][1-\text{E}[\theta]]}{\text{Var}(\theta)} - \left(1+\alpha_0+\beta_0\right) \end{aligned} \tag{11}$$

where α_0 and β_0 are the initial parameters assigned to $f_b(\theta|\alpha, \beta)$.

Using the model above, where $\alpha_0 = 7$, $\beta_0 = 3$, $\text{E}[\theta] = 0.56$, $\text{Var}(\theta) = 0.013$, and letting $n = 100,000$, we find that a noise-free sample n^* of merely **four** women would lead to the same posterior mean and standard deviation as in our model with noise. More generally, Table 2 shows the equivalent noise-free sample size that would

be needed to obtain the same posterior mean and variance for θ as in our model with noise, with $\theta \sim f_\beta(\theta|7,3)$ and $\lambda \sim f_\beta(\lambda|3,7)$, as n gets increasingly larger. In Appendix A, we derive the noise-free sample size as the sample goes to infinity (Theorem 1 from Appendix A enables us to compute the last row in the table).

Sample size (n)	Equivalent noise-free sample size for θ (n^*_θ)
10	1.4
100	3.2
$1,000$	3.6
$10,000$	3.7
$100,000$	3.7
$n \to \infty$	3.8

Table 2: Actual and equivalent noise-free sample sizes.

Note that in our example above, the equivalent noise-free sample size has an upper bound of 3.8. This reveals a drastic loss of information. This is because, as mentioned before, the likelihood function is unable to distinguish between numerous disparate explanations of the observed data (i.e., an infinite combination of θ and λ values lead to the same likelihood). Indeed, the equivalent noise-free sample size relative to the actual sample size remains disproportionately low even for tight distributions for λ or low expected values of λ. For example, Table 3 shows the equivalent noise-free sample size under different prior specifications for λ, keeping $n = 100,000$ fixed. Note that even if $\mathrm{E}[\lambda] = 0.0625$, a very low prior estimate of misclassification, n still shrinks to 117.

Noise estimate (λ)	Effective sample size
$\mathrm{E}[\lambda] = 0.2, \mathrm{Var}(\lambda) = 0.02$	$n^*_\theta = 4.7$
$\mathrm{E}[\lambda] = 0.1, \mathrm{Var}(\lambda) = 0.008$	$n^*_\theta = 16.3$
$\mathrm{E}[\lambda] = 0.0625, \mathrm{Var}(\lambda) = 0.003$	$n^*_\theta = 117.1$

Table 3: Equivalent noise-free sample sizes for different assumptions about λ.

In short, we do not have to bury our head in the sand, or overload our prior. By properly estimating the noise in the data, we typically avoid normative conflict. That is, by using the information we have regarding gender disparities in unemployment, historical

patterns of discrimination, implicit biases in the workplace, and so forth, and accepting the possibility of observed data coming from a noisy process as a result, one will ordinarily avoid normative conflict regardless of sample size in cases like Recruitment. While Recruitment is just an illustration, the potential for such misclassifications is a real risk in similar contexts involving hiring, lending, college admissions, and other domains where scarce resources are unevenly distributed across sensitive categories.

5 Discussion

In this section, we explain a little bit more generally how normative conflict need not arise under the model we develop, regardless of samples size, and we consider some broader discussion questions and areas for future work.

First, it is worth reflecting on the relationship between the model developed here and the original JKB approach. Our model is designed to complement and expand the latter, rather than to refute it. As mentioned earlier, when $\lambda = 0$ we recover the JKB model. And if one wishes, one can continue to rely on normative attitudes, as reflected in the epistemic risk function, in order to structure the prior in the model we develop. But the advantage of our approach is that while we may rely on normative attitudes to restructure the prior, we do not have to. Thus, unlike the JKB approach, our model is not vulnerable to the objection that we are avoiding normative conflict at the cost of our ability to learn.

Rather, we focus on capturing the noise in the data generating process. And we show that once one captures that noise reasonably well, the ensuing rational belief will not necessarily be one that leads to normative conflict, regardless of the difference in observed proportions between, say, talented men and talented women. Another way to put the point is that the conflict is avoided because uncertainty about both θ and λ does not completely go away even with an infinite sample size (as we show in Appendix A).

Thus, our approach encourages learning by incorporating the full information about the relevant situation. In the case of Alice and the Recruitment example, that includes the observed proportion of women who are deemed talented, but it also includes everything else we know about gender disparities in the workforce, historical patterns of discrimination, implicit attitudes and bias, and systemic institutional barriers that may undermine women's success. The model encourages us to include what we know by setting up a prior for θ, for λ, and, as we will see in Appendix B, for a parameter, ξ, reflecting errors in the other direction as well (untalented women who are recorded as talented).

It is also worth highlighting that in our approach, we do not avoid normative conflicts a priori, so to speak. That is, we do not write out the possibility of normative conflict as a mathematical fact. Rather, we avoid normative conflict because the information about the workforce that is reflected in our prior reflects the world we live in – i.e., a world where women have been historically discriminated against. It is of course possible to assume the opposite – that women are favored in the workforce – and then epistemic rationality would require one to believe that women like Alice are less talented then their male counterparts. But this is not a problem because in a possible world where women dominate the workforce and are privy to the privileges that men are provided in our world, gender stereotypes would no longer remain the same either. Indeed, in such a hypothetical world, we might worry about the opposite – statistical information supporting stereotypes reinforcing men's talent or aptitude. And using our model, we would then structure the priors on θ, λ, and ξ to capture the noise with respect to men. The same general point can be made about any group distinction or number of groups.

Second, one might worry that by using the noise model, we could amplify the notion that a disadvantaged group doesn't possess some of the characteristics that are necessary to be successful or to be *deemed* talented and thus create further normative conflicts. Consider en example. Suppose that one of the unobservable and/or unmeasurable attributes to be deemed talented in an industry like banking is membership in an "old boys club." Then it is of course possible that our noise model would amplify the absence of this attribute among women – i.e., the absence of women in old boys' clubs. While this is true, the belief that women are more likely to be absent in old boys' clubs is not a further normative conflict. Instead, it is evidence of just the kinds of injustice that lead to the disparity in the proportion of women who are deemed talented as compared to men to begin with. In other words, normative conflict arises when we believe that women are overall less likely to be talented, and not when we believe that they are overall less likely to be members of old boys' clubs. The latter merely reflects the unfair advantage of the advantaged group. Indeed, it is information we should not ignore if we want to mitigate gender disparities in the workforce.

Third, it is worth emphasizing that we do not intend to banish, so to speak, all normative attitudes from the identification of an appropriate model. We agree with JKB that in the absence of any information about a parameter, the choice of prior reflects, to some extent, normative attitudes to the cost of error. What we would like to avoid, though, is situations where normative attitudes overwhelm inference in such a way that the agent appears to be incapable of reasonably responding to evidence. And the cases where the JKB account is most vulnerable are cases where the sample size is very large, such as Recruitment. By introducing the possibility of misclassification,

we show that epistemic rationality does not necessarily require a stereotypical belief, regardless of sample size, and we do so without tinkering with the resilience, as Joyce ([2005]) calls it, of the prior. The key message that we would like to drive home here is that a reasonable model for cases like Recruitment needs to include uncertainty about λ, but a precise estimate of λ is not needed and in practice not realistic. For example, one may start with a uniform prior for λ due to background assumptions that mistakes in either the false positive or false negative error direction are equally bad. This would exert only a negligible effect on inference after we update λ on thousands of observations, as in Recruitment.

Fourth, we would like to further highlight why the massive loss of information occurs. As we illustrate in Table 1, under a modest misclassification rate the effective sample size remains extremely low even as data gets arbitrarily large. And in Theorem 1 in Appendix A, we reinforce this point by highlighting that the effective sample size is asymptotically bounded by 3.8. One might wonder whether we are baking this conclusion in by specifying too high of a misclassification rate. But we are not. In Table 2, we present a number of different models, and in each row we specified a different prior for λ, with increasingly less misclassification in descending order. And regardless of how small the error rate gets, the loss of information remains overwhelming. For example, even when the misclassification rate is as low as 6%, and the variance around that estimate extremely tight (last row of Table 2), the effective sample size still shrinks from $100,000$ to just over 100.

To understand why this occurs, it is helpful to recall that we can think of the posterior as a weighted mixture of many posteriors, one for each possible misclassification (Equation 8). That is, the posterior is a mixture of the posterior we would have if only one woman is misclassified, multiplied by the probability that only one woman is misclassified, plus the posterior one would have if exactly two women were misclassified, multiplied by the probability that exactly two women were misclassified, etc. The information loss occurs even with a low misclassification rate because there are so many possible ways the true state of nature could turn out to be. Of course, as the misclassification rate gets asymptotically close to 0, the effective sample size will not shrink as much. But the highlight is not so much that effective sample size collapses for any non-zero misclassification. Rather, it is just how sensitive the effective sample size, n^*, is to prior uncertainty about λ, and how quickly it diminishes, given very modest assumptions about noise.

6 Concluding Remarks

Consider again the passage from Gendler that we started with: in an unequal society one can not be simultaneously moral and epistemically rational. If we know that more crimes in our society are on average committed by a minority group, the thought goes, then epistemic rationality seems to force us to use that frequency in estimating the future criminality of any particular member of that group – in a way that seems patently unjust. But if we know our society is unjust, then we also know that the disparity in observed crime rates is occurring because of, for example, bias against this group, over policing of its minority neighborhoods, harsh prosecution of petty crimes, etc. And our model encourages us to take this information into account as we go from the observed disparity in crime rates across groups to the prediction that any particular individual from the minority group will commit a crime in the future. When we do this, epistemic rationality is no longer misaligned with considerations of justice. The same can be said for performance differences in academia, finance, and so forth, across many different protected classes, including race, ethnicity, and gender. Thus, in most cases we are likely to encounter, normative conflicts are avoidable. But the requirement of noise does mean that we cannot guarantee they will be avoided. The absence of such a guarantee is a virtue, rather than a vice. It means that when there are true underlying population differences – differences which must be addressed in order to reduce the kinds of socioeconomic disparities that give rise to unequal statistics in the first place – our model will enable us to detect them.

Appendix A

In this Appendix we derive the posterior distribution for θ when $n \to \infty$ with the proportion of women who are deemed talented, $\varphi_0 = \gamma/n$, being held constant. We will exploit the fact that as $n \to \infty$, the likelihood (7) converges to the Dirac delta function $\delta\left(\theta\left(1-\lambda\right)-\varphi_0\right)$.

Theorem 1. Let the prior distribution for (θ, λ) be a product of independent beta distributions $f_\beta(\theta|a_\theta, b_\theta)$ and $f_\beta(\lambda|a_\lambda, b_\lambda)$, and let the likelihood be the Dirac delta function $\delta\left(\theta\left(1-\lambda\right)-\varphi_0\right)$. Then, up to normalization, the marginal posterior pdf for θ is

$$f\left(\theta|\varphi_0\right) \propto \begin{cases} \theta^{a_\theta - a_\lambda - b_\lambda}\left(1-\theta\right)^{b_\theta - 1}\left(\theta-\varphi_0\right)^{a_\lambda - 1} & \text{if } \theta \geq \varphi_0, \\ 0 \text{ if } \theta < \varphi_0. \end{cases}$$

Proof. When $n \to \infty$, the joint posterior pdf is

$$f(\theta, \lambda|\varphi_0) \propto f_\beta(\theta|a_\theta, b_\theta) f_\beta(\lambda|a_\lambda, b_\lambda) \delta\left(\theta\left(1-\lambda\right) - \varphi_0\right),$$

and the marginal posterior for θ is

$$f\left(\theta|\varphi_0\right) \propto f_\beta(\theta|a_\theta, b_\theta) \left(\int_0^1 f_\beta(\lambda|a_\lambda, b_\lambda) \delta\left(\theta\left(1-\lambda\right) - \varphi_0\right) d\lambda \right).$$

Denote $\varphi = \theta\left(1-\lambda\right)$; then $\lambda = 1 - \varphi/\theta$, $d\lambda = -\frac{1}{\theta} d\varphi$, and

$$\begin{aligned}
&\int_0^1 f_\beta(\lambda|a_\lambda, b_\lambda) \delta\left(\theta\left(1-\lambda\right) - \varphi_0\right) d\lambda \\
&= \int_0^\theta f_\beta(1-\varphi/\theta|a_\lambda, b_\lambda) \delta\left(\varphi - \varphi_0\right) \frac{1}{\theta} d\varphi \\
&= \begin{cases} \frac{1}{\theta} f_\beta\left(1-\varphi_0/\theta|a_\lambda, b_\lambda\right) & \text{if } \theta \geq \varphi_0, \\ 0 \text{ if } \theta < \varphi_0. \end{cases}
\end{aligned}$$

Therefore,

$$f\left(\theta|\varphi_0\right) \propto \begin{cases} \frac{1}{\theta} f_\beta(\theta|a_\theta, b_\theta) f_\beta\left(1-\varphi_0/\theta|a_\lambda, b_\lambda\right) & \text{if } \theta \geq \varphi_0, \\ 0 \text{ if } \theta < \varphi_0. \end{cases}$$

In turn, $f_\beta(\theta|a_\theta, b_\theta) \propto \theta^{a_\theta - 1}\left(1-\theta\right)^{b_\theta - 1}$ and $f_\beta(\lambda|a_\lambda, b_\lambda) \propto \theta^{a_\lambda - 1}\left(1-\theta\right)^{b_\lambda - 1}$, so

$$\begin{aligned}
f\left(\theta|\varphi_0\right) &\propto \begin{cases} \frac{1}{\theta} \theta^{a_\theta - 1}\left(1-\theta\right)^{b_\theta - 1}\left(1-\varphi_0/\theta\right)^{a_\lambda - 1}\left(\varphi_0/\theta\right)^{b_\lambda - 1} & \text{if } \theta \geq \varphi_0, \\ 0 \text{ if } \theta < \varphi_0 \end{cases} \\
&\propto \begin{cases} \theta^{a_\theta - a_\lambda - b_\lambda}\left(1-\theta\right)^{b_\theta - 1}\left(\theta-\varphi_0\right)^{a_\lambda - 1} & \text{if } \theta \geq \varphi_0, \\ 0 \text{ if } \theta < \varphi_0. \end{cases}
\end{aligned}$$

□

The marginal posterior pdf for λ is derived similarly, and is given by

$$f\left(\lambda|\varphi_0\right) \propto \begin{cases} \lambda^{a_\lambda - 1}\left(1-\lambda\right)^{b_\lambda - a_\theta - b_\theta}\left(1-\varphi_0-\lambda\right)^{b_\theta - 1} & \text{if } \lambda \leq 1-\varphi_0, \\ 0 \text{ if } \lambda > 1-\varphi_0. \end{cases}$$

To compute the equivalent sample size as in Table 1, we first find parameters $a_\theta^*$and b_θ^* of the beta distribution that match the first two moments of $f\left(\theta|\varphi_0\right)$. Then the

equivalent sample size equals $n_\theta^* = a_\theta^* + b_\theta^* - a_\theta - b_\theta$. From the equations $E(\theta|\varphi_0) = \frac{a_\theta^*}{a_\theta^* + b_\theta^*}$ and $V(\theta|\varphi_0) = \frac{a_\theta^* b_\theta^*}{\left(a_\theta^* + b_\theta^*\right)^2 \left(a_\theta^* + b_\theta^* + 1\right)}$ we find $a_\theta^* + b_\theta^* = \frac{E(\theta|\varphi_0)(1-E(\theta|\varphi_0))}{V(\theta|\varphi_0)} - 1$. In particular, for $a_\theta = 7$, $b_\theta = 3$, $a_\lambda = 3$, $b_\lambda = 7$, and $\varphi_0 = 0.3$ we get $n_\theta^* = 3.8$.

Appendix B

In this Appendix, we expand the model to accommodate two-way errors. That is, suppose we have misclassification in both directions so that, using our Recruitment example, actually talented women are sometimes recorded as untalented, with rate λ, and actually untalented women are sometimes recorded as talented, with rate ξ. Graphically, the situation now looks like this.

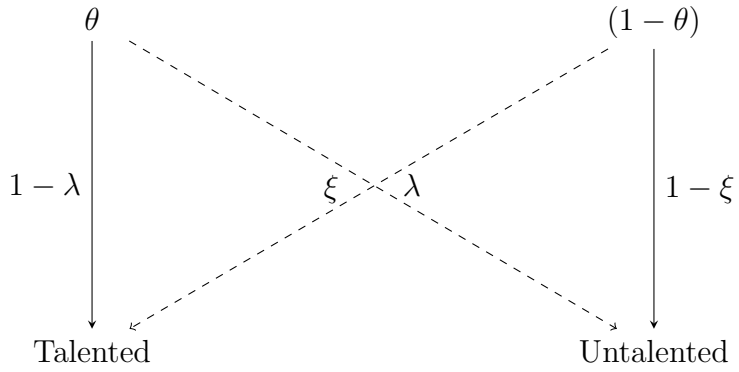

Figure 5: Recruitment with noise levels λ and ξ

The probability that a woman is classified as talented is now $\phi^* = \theta(1-\lambda) + (1-\theta)\xi$. And the probability that a woman is classified as untalented is $(1-\theta)(1-\xi) + \theta\lambda$. The likelihood is now Bernoulli in ϕ^*,

$$f(y|\theta, \lambda, \xi) = \phi^{*y}[1-\phi^*]^{n-y}.$$

Using a beta prior for both noise parameters, the full prior is,

$$\pi(\theta, \lambda, \xi) \propto \pi(\theta|\alpha_\theta, \beta_\theta)\pi(\lambda|\alpha_\lambda, \beta_\lambda)\pi(\xi|\alpha_\xi, \beta_\xi).$$

The posterior is

$$\pi(\theta, \xi, \lambda|y) \propto \pi(\theta, \xi, \lambda) f(y|\theta, \xi, \lambda).$$

Finally, the marginal posterior distribution for θ is

$$\pi(\theta|y) \propto \int_0^1 \int_0^1 \pi(\theta, \xi, \lambda|y) d\lambda d\xi.$$

Using the same Monte Carlo simulations as in the main text, but with a now weakly informative beta prior on ξ, we find that $n^* < 2$. This is to be expected because we have now increased the number of ways in which different values can be assigned to these parameters.

Acknowledgments

Thank you to two anonymous reviewers from *The British Journal for the Philosophy of Science*, whose comments and suggestions have undoubtedly improved the final version of this paper. Thanks also to Sina Fazelpour, Zoë Johnson-King, James Joyce, and Snow Xueyin Zhang for their very helpful comments, and to seminar participants at the University of Toronto, University of Washington, and the 2020 Formal Epistemology Workshop (UC Irvine).

Boris Babic
Department of Decision Sciences
INSEAD
France and Singapore
boris.babic@insead.edu

Anil Gaba
Department of Decision Sciences
INSEAD
France and Singapore
anil.gaba@insead.edu

Ilia Tsetlin
Department of Decision Sciences
INSEAD
France and Singapore
ilia.tsetlin@insead.edu

Robert L. Winkler
Department of Decision Sciences
Duke University,
Fuqua School of Business
Durham, NC
rwinkler@duke.edu

References

Babic, B. [2019]: 'A Theory of Epistemic Risk', *Philosophy of Science*, **86**(3), pp. 522–50.

Basu, R. [2018]: 'The Specter of Normative Conflict: Does Fairness Require Inaccuracy?', in E. Beeghly and A. Madva (eds), *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*, London: Routledge.

Basu, R. and Schroeder, M. [2019]: 'Doxastic Wronging', in B. Kim and M. McGrath (eds), *Pragmatic Encroachment in Epistemology*, London: Routledge.

Gaba, A. [1993]: 'Inference with an Unknown Noise Level in a Bernoulli Process', *Management Science*, **39**(10), pp. 1179–97.

Gaba, A. and Winkler, R.L. [1992]: 'Implications of Errors in Survey Data: A Bayesian Model', *Management Science*, **38**(7), pp. 913–25.

Geman, S. and Geman, D. [1984]: 'Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(6), pp. 721–41.

Gendler, T.S. [2011]: 'On the Epistemic Costs of Implicit Bias', *Philosophical Studies*, **156**(1), pp. 33–63.

Greaves, H. and Wallace, D. [2006]: 'Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility', *Mind*, **115**(459), pp. 607–32.

Howson, C. and Franklin, A. [1994]: 'Bayesian Conditionalization and Probability Kinematics', *British Journal for the Philosophy of Science*, **45**(2), pp. 451–66.

Huttegger, S. [2017]: *The Probabilistic Foundations of Rational Learning*, Cambridge University Press.

Pettigrew, R. [2016]: *Accuracy and the Laws of Credence*, Oxford University Press.

Johnson-King, Z. and Babic, B. [2020]: 'Moral Obligation and Epistemic Risk', in M. Timmons (ed), *Oxford Studies in Normative Ethics*, Vol. 10, Oxford University Press.

Joyce, J.M. [2005]: 'How Probabilities Reflect Evidence', *Philosophical Perspectives*, **19**(1), pp. 153–78.

Joyce, J.M. [1998]: 'A Nonpragmatic Vindication of Probabilism', *Philosophy of Science*, **65**(4), pp. 575–603.

Joyce, J.M. [2009]: 'Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief', in F. Huber and C. Schmidt-Petri (eds), *Degrees of Belief*, Springer.

Lindley, D.V and Phillips, L. [1976]: 'Inference for a Bernoulli Process (A Bayesian View)', *The American Statistician*, **30**(3), pp. 112-9.

Plummer, M. [2003]: *JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling*.

van Fraassen, B.C. [1981]: 'A Problem for Relative Information Minimizers in Probability Kinematics', *British Journal for the Philosophy of Science*, **32**(4), pp. 375–79.

Winkler, R.L. and Gaba, A. [1990]: 'Inference with Imperfect Sampling from a Bernoulli Process', in Geisser, S.P.S, Hodges, J.S. and Zellner, A. (eds), *Bayesian and Likelihood Methods in Statistics and Econometrics*, North-Holland.