Norms, Stereotypes, and Accuracy

Boris Babic

California Institute of Technology

Abstract

Philosophers have been substantially interested in statistically warranted stereotypes. We often find ourselves in situations where we are aware of a statistical generalization pertaining to a relatively immutable physical characteristic and we need to formulate a credence about a particular individual possessing that characteristic. In such cases, it seems reasonable to assume that our credence about the individual should correspond to the known frequency. However, if the frequency tracks an ethnic, racial, or gender class, adopting a credence about the individual on this basis appears to be unjust. This suggests an apparent normative conflict between the requirements of epistemic rationality and the obligations of ordinary morality. In this paper, we argue that by properly taking into account an agent's attitudes to epistemic risk, such normative conflicts can be avoided.

1 Introduction

How should we respond to statistical evidence supporting stereotypes based on relatively immutable physical characteristics? For example, you learn about a study purporting to have found a statistically significant (and strong) relationship between a racial, gender, or ethnic group and the commission of certain types of crime. You then come across an individual who belongs to that group. Should your credence about this person's propensity to commit that crime correspond to the frequency (sample mean) reported in the study? At first blush, the answer appears to be 'yes'. In the epistemology literature, many authors have defended a *frequency-credence principle* which requires an agent to identify their credence with the reported frequency (White, 2010) (italicized terms will be defined). However, it has been argued that holding such prejudicial beliefs, even if they are statistically warranted, is wrongful (Basu and Schroeder, 2018). Further, even if one denies that we can wrong others by our beliefs about them, acting on the basis of such beliefs can certainly be unjust, as Mogensen (2018), Buchak (2014) and others argue. This suggests that if the frequency-credence principle indeed holds in cases involving racial, ethnic or gender reference classes, then the norms of epistemic rationality will often be incompatible with the obligations of ordinary morality. In such cases, it appears impossible to be both rational and just. Gendler (2011) refers to the ubiquity of conflict between our epistemic requirements and moral obligations in a world characterized by stark socioeconomic inequality as the ‘sad conclusion’.

In this paper, we argue that in the accuracy based approach to epistemology, Gendler’s sad conclusion can ordinarily be avoided. To do this, we defend a theory of epistemic rationality according to which an agent’s attitudes to *epistemic risk* affect the credences she holds. Levinstein (2017) makes a similar suggestion. However, Levinstein does not make clear what epistemic risk is, how attitudes to epistemic risk affect an agent’s credences, or how to measure epistemic risk attitudes. We make this explicit by relying on a theory of epistemic risk articulated in Babic (2017). We then apply the theory to the case of statistically warranted stereotypes giving rise to *normative conflict*. We show that ideal epistemic rationality is compatible with many different attitudes to epistemic risk, that the frequency-credence principle does not apply in cases that would ordinarily give rise to normative conflict, and therefore that in most cases involving statistically warranted stereotypes, there are permissible attitudes to epistemic risk that would give rise to credences which are not wrongful.

The paper proceeds as follows. In Section 2, we provide a simple example of normative conflict and define, following White (2010), the frequency-credence principle. In Section 3, we give a directed introduction to accuracy based epistemic rationality. As applied to cases of interest to us, epistemic rationality requires that we adopt credences which maximize expected epistemic utility where epistemic utility is given by an appropriate accuracy measure or *scoring rule*. This imposes several coherence and symmetry constraints on an agent’s credences, which we will make explicit. In Section 4, we introduce a distinction between direct inference and predictive inference. Most cases giving rise to normative conflict require a predictive inference. In such cases, an agent’s prediction need not conform to the frequency. Instead, it must be made in a way that satisfies what Huttegger (2017) calls the *generalized rule of succession*. The generalized rule of succession is an inference schema rather than a particular rule. In Section 5, we introduce the notion of epistemic risk, as developed in Babic (2017). We then use the epistemic risk framework to explain how normative attitudes to error can affect the shape of an agent’s scoring rule. In section 6, we explain how this framework enables us to avoid normative conflicts. In particular, an agent’s attitudes to how much epistemic risk they are willing to assume, together with her scoring rule, determine her specification of the generalized rule of succession. While some specifications would lead to predictions giving rise to normative conflict, other equally permissible specifications would not. Section 7 concludes.

2 Normative Conflict

Rational belief, it is said, aims at truth, and rational credences aim at accuracy (Railton, 1994; Wedgwood, 2002). As a result, the norms governing belief and credence appear to be separable from the norms of ordinary morality – such as justice, desert, or fairness. Therefore, it seems prima facie possible to find oneself in a dilemma between competing normative requirements – that is, between the requirements of epistemic rationality and the obligations of ordinary morality. This situation has been described as normative conflict (Basu, 2018). We adopt the following rough definition.

> **Normative Conflict**. An agent is under Normative Conflict if *both*, from an epistemic perspective, the agent ought to believe/be very confident in a proposition ϕ *and*, from a moral perspective, the agent ought not believe/be as confident in ϕ.

In the next section, we comment further on how beliefs might be wrongful. Assuming they can be, as many authors have argued, we show that in the accuracy-first framework of epistemic rationality, normative conflicts can usually be avoided. The status of certain special cases, which we will identify, is less clear.

2.1 Frequencies and Stereotypes

Consider the following example.

> **Gender Bias Study**. One morning you read in the Washington Post about a study reporting gender discrepancies in academic employment in the United States. The authors of the study surveyed 100 people, 50 men and 50 women. They found that 70% of women held administrative or clerical roles and 30% held faculty roles while the opposite was true of men.

The results of the Gender Bias Study are summarized in the following table.

	Male	Female
Admin	15	35
Faculty	35	15

Table 1: Gender Bias Study

Since we are interested in assessing differences between groups using binary data, it would be common to see the p-value computed under the null hypothesis of no gender effect using Fisher's Exact Test. Using the data from Table 1, $p < 0.0001$. The results are statistically significant. Suppose further that the news report is accurate and that the study was pre-registered and conducted competently. For example, the authors did not perform multiple comparisons of the data they collected in order to find evidence of gender bias, it was stipulated in advance that the survey would end after 50 men and 50 women responded, they disclosed all covariates that were tracked, and no observations were excluded. While no study is perfect, there are no red flags. The sample size in this example is quite generous as compared to ordinary research in highly regarded psychology journals. Holmes et al. (2011) report the median sample sizes over the last 30 years in four leading journals of the American Psychological Association. In each of the time periods studied, and across all journals, the median group size is significantly less than 50 (Table 3 in Holmes et al. (2011)).

We have chosen an example that perhaps reinforces certain gender stereotypes but it is easy to imagine similar studies reporting disparities in the commission of crimes, aptitude or intelligence. In general, the kinds of cases we are interested in are cases Tetlock et al. (2000) call 'taboo base-rates': frequencies that track morally sensitive categories like race, gender and ethnicity in a way that reinforces societal stereotypes about such groups.

2.2 The Frequency-Credence Connection

Suppose after learning about the Gender Bias Study you encounter Mary, a woman employed at your local university. Suppose further that before reading about the Gender Bias Study, you had no information about the relative distribution of men and women across different occupational roles in academia. What should your credence be in the proposition 'Mary is a faculty member'? A natural answer is .3 (see Buchak, 2014, for example). This answer is justified on the basis of the following principle, as given in White (2010).

> **Frequency-Credence Principle**. If, (a) I know that a is an F, and (b) I know that $freq(G|F) = x$, and (c) I have no further evidence bearing on whether a is a G, then, $Cr(a \text{ is a } G) = x$.

An early statement of the Frequency-Credence Principle may be found in Reichenbach (1938, 1949). The Principle was forcefully defended by Kyburg (1974) while Levi (1977) puts some important constraints on its scope of application, which we will note. White (2010) illustrates the Principle as follows.

> [Suppose] you somehow learn that 37% of formal epistemologists are left-handed. With *only* this to go on, what should your credence be that Branden Fitelson is left-handed? The very natural answer is 37% (pg. 169, emphasis in original).

When we apply the Frequency-Credence Principle to cases involving taboo reference classes, as Tetlock et al. (2000) call them, we often end up in a position of Normative Conflict.

From the perspective of epistemic rationality, it appears that since we do not know anything about Mary, and we had no prior information about women in the academic workplace, we should be very confident, in particular our credence should be .7, that Mary holds an administrative position. From a moral perspective, however, many people share the judgment that it would be unfair to Mary to assume she holds an administrative role at the university. There are two ways that a high credence about Mary's occupation may be wrongful. First, you may take some action on the basis of your credence with respect to Mary. For example, you learn that an administrative person caused something bad to happen. You have two suspects, Mary and Morty. You do not have any other information about these individuals. Since Mary is more likely than Morty to hold an administrative role, you blame Mary.[1] Second, Basu and Schroeder (2018) suggest that it is possible to wrong someone by your beliefs about them (even if you do not act on these beliefs). In this case, it may be wrongful toward

[1] Buchak (2014) argues that because high credences can support morally wrongful acts, such as blaming Mary in this example, they are not appropriate doxastic states for reactive attitudes like blame or praise. Instead, we need a different doxastic state; namely, belief. Weatherson (2014) agrees that blame is inappropriate but resists Buchak's conclusion that a different doxastic state is called for in cases like this. We will see that this conclusion is avoidable once we take into account an agent's attitudes to epistemic risk.

Mary to be so confident that she is not a faculty member on the basis of statistical evidence about a class to which she just so happens to belong.

We will assume for the purpose of this project that a credence about Mary based on a frequency such as the one reported in the Gender Bias Study may be morally wrongful for either (or both) of these reasons. Therefore, it appears that a rational Bayesian agent should be confident that Mary is not a faculty member whereas a morally good person would be more cautious. As a result, becoming aware of the Gender Bias Study together with the Frequency-Credence Principle puts you in a position of Normative Conflict. Since we live in a society characterized by stark ethnic, racial, and gender disparities, normative conflicts are very common. Gendler (2011) calls the ubiquity of normative conflicts the 'sad conclusion' which she describes 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, towards 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* (pg. 57, emphasis added).

3 Epistemic Rationality

In this section, we provide a simplified introduction to accuracy based epistemic rationality. Suppose we have a hypothesis of interest, h. Using our Gender Bias Study let,

$$h := \text{`Mary holds an administrative position'}$$

Our space of hypotheses, denoted by Ω, is either that Mary is a faculty member or that Mary is an admin (when we get to the section on predictive inference, we will use a random variable to identify these outcomes). In other words, h and its negation $\overline{h}$ create a partition of Ω. Let $\mathcal{F}$ be a set consisting of subsets of Ω formed by taking finite unions and intersections which is closed under complementation and and includes the null set. $\mathcal{F}$ represents the event space. We will say that $\mathcal{F}$ is an *algebra* (for countable sets, it is a σ-algebra). Finally, let $\mathcal{P}$ be the set of probability distributions on $\mathcal{F}$. A function is a probability distribution if it satisfies the Kolmogoroff axioms.[2] The collection $(\Omega, \mathcal{F}, \mathcal{P})$ is a probability space.

Our decision-makers prefer more accurate credences to less accurate ones. But how should we measure inaccuracy?[3] It is natural to suppose that if h is true, the higher the credence in h, $pr(h)$, the lower one's inaccuracy score, and if h is false, the lower the credence in h the lower one's inaccuracy score.[4] That is, inaccuracy should be a *truth-directed* monotonic function of the probability assigned to h. This reflects in graded terms the Jamesian

[2] This means, for finite event spaces, as those that will be of interest to us, that for $e \in \mathcal{F}$, $Pr(e) \geq 0$, $\sum_{\mathcal{F}} e_i = 1$, and $Pr(\bigcup_{\mathcal{F}} e_i) = \sum_{\mathcal{F}} Pr(e_i)$.

[3] It is common to use *in*accuracy with 0 as the bound on how well an agent can do.

[4] We use lowercase p, q, ... to identify a probability measure defined over the event space $\mathcal{F}$ and we use the shorthand $pr(h)$ to indicate the probability assigned to the atom or hypothesis h in Ω. For coherent agents, whose credence function is a probability, if we know the distribution on Ω we know the distribution on $\mathcal{F}$.

dictum that in epistemology we ought to seek truth and avoid error. It is also reasonable to assume that regardless of the shape of our measure of inaccuracy, it should be *continuous* so as to avoid arbitrarily small changes in credence leading to large changes in inaccuracy.

Inaccuracy is typically denoted by a two-place function $s : [0,1] \times \{0,1\} \to \mathbb{R}$, that takes the true value of h (denoted by '0' if false and '1' if true), and the probability assigned to h, and gives us a real number, where lower numbers are better. We will denote this function $s_v(pr(h))$, where v is a binary indicator for the event. This is often called a *scoring rule*. For example, a common scoring rule is squared Euclidean distance. If h is true, the score would be $s_1(pr(h)) = (1 - pr(h))^2$ whereas if h is false it would be $s_0(pr(h)) = (0 - pr(h))^2$. This is also known as the Brier score, named after the meteorologist Glenn Brier who first proposed it as a measure of accuracy for probabilistic weather forecasts. Ordinarily, our scoring rules are additive, so that we measure inaccuracy by taking the sum of inaccuracies of every event in the partition of interest to us. This condition is not essential, however. Since our decision-makers do not know whether h is true or false in advance, their goal is to assign credences in a way that minimizes *expected* inaccuracy. That is, they ought to choose p to minimize $\mathrm{E}_p[s_v(p)]$. If their measure of inaccuracy satisfies the following property, we say it is *proper* (if the inequality is strict, it is *strictly proper*).

$$\text{For all } p, q \in \mathcal{P},\ \mathrm{E}_p[s_v(p)] \leq \mathrm{E}_p[s_v(q)]$$

This property says that from the agent's perspective her own credences are at least as good in expectation as any other credences she could adopt.

If a scoring rule, which we will understand here as the agent's epistemic utility, satisfies (1) truth-directedness, (2) continuity and (3) strict propriety, then we can derive the following important epistemic norms. First, an epistemically rational agent ought to be coherent. That is, their credences can be represented by $p \in \mathcal{P}$, a function that assigns a probability to every event in $\mathcal{F}$ defined with respect to Ω. This is because for every incoherent credence function there exists an accuracy dominating coherent one (Joyce, 1998, 2009). Coherence, therefore, is justified through the decision-theoretic principle *dominance* as applied to a measure of inaccuracy satisfying (1)-(3). Second, when agents receive evidence which constitutes a partition of Ω, they should update their credences by Bayesian conditionalization (Greaves and Wallace, 2006; Easwaran, 2013). That is, if they learn a their new credence function should be $pr^*(h) = pr(h|a) := pr(h \cap a)/pr(a)$ provided $pr(a) \neq 0$. This update rule minimizes the prior expected inaccuracy of their posterior credences.[5] Moreover, even if the evidence is not a partition of Ω, conditionalization minimizes expected inaccuracy provided the outcomes constitute a σ-algebra (Huttegger, 2013).[6] Therefore, Bayesian updating is justified using the decision-theoretic principle *minimize expected inaccuracy* where inaccuracy is given by a scoring rule satisfying (1)-(3).

We will assume that our agents are epistemically rational. This means their measure of epistemic utility satisfies (1)-(3) and as a result they start with a coherent prior credence

[5] Leitgeb and Pettigrew (2010) derive a similar result using the quadratic score in particular.

[6] Huttegger (2014) examines updating in more general learning situations where we cannot update by simple Bayesian conditionalization. Following, van Fraassen (1995), he defends a reflection principle between prior and posterior probabilities known as the martingale condition, of which conditionalization is a specific instance. The cases of interest to us will be cases where ordinary Bayesian updating is possible.

function which they update by Bayesian conditionalization. We *want* to make these assumptions in order to evaluate whether such an agent will be forced by these obligations into a position of normative conflict. In the next section, we introduce one more principle of epistemic rationality, which is an application of coherence and updating norms to the specific context of predictive inference. Huttegger (2017) calls it the generalized rule of succession. This rule will enable us to evaluate the Frequency-Credence Principle as applied to Mary.

4 Direct and Predictive Inference

Recall the Frequency-Credence Principle: if we learn the proportion of F's that are G's is x, and a is an F, then, assuming no other information about a, our credence that a is a G should be x. As stated, this principle is ambiguous. It is unclear whether a was included in the sample from which the frequency was reported.

If a was part of the sample used to determine the frequency, then the inference to be made is often called direct inference or statistical syllogism (Kyburg, 1974; Levi, 1977). We are reasoning down from knowledge about all F's to a particular a that is an F. In this case, as Levi (1977) argues, whether or not our credence should match the frequency depends on whether a was randomly selected for presentation. A better way of putting this point is to say that whether or not the credence should match the frequency depends on whether the prior probabilities assigned to every member in the sample are exchangeable (we define exchangeability, below). If they are then, plausibly, the frequency data swamps any prior beliefs we might have had and the credence should match the frequency.

However, most frequency data giving rise to Normative Conflict are not like this. Consider the Gender Bias Study again. Instead of reasoning down, we are reasoning laterally, so to speak, in order to predict the value of a new data point which has not yet been observed. This is sometimes called inverse inference, but this expression is likewise ambiguous, as it can mean either an inference about an unknown parameter of interest, or a prediction about an unobserved data point based on a posterior distribution or estimate of that unknown parameter. The latter case is of interest to us, and it is ordinarily called predictive inference (Gelman et al., 2013).

Consider the Gender Bias Study. To make the problem precise, let X_{ij} be a random variable where $i \in \{1, 2\}$ denotes whether person $j \in n$ is male or female, respectively, $X_{ij} = 0$ if j is a faculty member, and $X_{ij} = 1$ if j is an admin. For illustration, the results of the study might be tabulated as follows.

j	X_{1j}	X_{2j}
1	0	1
2	0	1
3	0	1
4	1	0
5	1	1
...	...	...
n	0	1

Table 2: Sample coding of covariates in Gender Bias Study

We will use lowercase x_{ij} for observed values of X_{ij}. To establish a norm for predictive inference, we introduce two more concepts: exchangeability and conditional independence.

Exchangeability. Let $p(x_{i1}, ..., x_{in})$ be the joint distribution of $X_{i1}, ..., X_{in}$. If the following is true,

$$p(x_{i1}, ..., x_{in}) = p(x_{\sigma(i1)}, ..., x_{\sigma(in)})$$

for each permutation σ of the integers from 1 to n then the sequence $X_{i1}, ..., X_{in}$ is said to be exchangeable.

Exchangeability is a property of the agent's credences. It says that they are invariant to reordering the individuals in the sample or that the subscript label does not convey any information.

Conditional Independence. Let

$$\hat{\theta}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij}$$

be the sample mean for group i. If the following is true,

$$p(x_{ij} | \hat{\theta}_i, x_k, k \neq j) = \hat{\theta}_i$$

Then $X_{i1}, ..., X_{in}$ are conditionally independent.

De Finetti (1974)'s celebrated representation theorem implies that if $X_{i1}, ..., X_{in}$ are exchangeable then they are conditionally independent.

In the Gender Bias Study, your prior beliefs about the individual men and individual women are exchangeable within each group. The only information you have about these survey respondents are the summary statistics reported in the study. Therefore, absent any other information, we may treat them as conditionally independent given $\hat{\theta}_i$. As a result, the probability that any given man in the sample is an admin is given by $\hat{\theta}_1 = 0.3$ and the probability that any given woman in the sample is an admin is given by $\hat{\theta}_2 = 0.7$. However, the inference about Mary is not a direct inference. Using our two-sample notation, it is a predictive inference about $X_{2(n+1)}$ based on a frequency observed in the sample $x_{21}, ..., x_{2n}$. How should we make such an inference?

Let θ_i be the (unknown) *population* mean among group i. Let $p(\theta_i)$ be the prior probability distribution for θ_i. Further, let $f(x_i|\theta_i)$ be the sampling distribution of a single random variable x_i in group i. Since each X_i is categorical with two possible outcomes, $f(x_i|\theta_i)$ follows a binomial distribution.[7] A flexible parametric prior distribution for θ_i is the beta distribution $p(\theta_i|\alpha, \beta)$ with two shape parameters $\alpha > 0$ and $\beta > 0$, whose density is given by,

$$p(\theta_i) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}$$

where $\Gamma(n) = (n-1)!$. Since the beta distribution is known to be conjugate to the binomial distribution,[8] the posterior distribution for θ_i after observing $x_{i1}, ..., x_{in}$ will likewise be a beta distribution whose density is given by,

$$p(\theta_i|x_{i1}, ..., x_{in}) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\hat{\theta}_i+\alpha)\Gamma(n(1-\hat{\theta}_i)+\beta)}\theta_i^{n\hat{\theta}_i+\alpha-1}(1-\theta_i)^{n(1-\hat{\theta}_i)+\beta-1}$$

The beta-binomial distribution lends itself to a very intuitive interpretation. We know that $n\hat{\theta}_i$ is the sum of admins in group i and $1 - n\hat{\theta}_i$ is the sum of faculty in group i. In this case, α is the number of 'pseudo' observations of admins and β is the number of 'pseudo' observations of faculty. If $\alpha = \beta = 1$ the prior distribution for the unknown rate θ_i is uniform. The posterior distribution, i.e., the distribution of θ_i after updating by Bayesian conditionalization on the evidence $x_{i1}, ..., x_{in}$, is given by adding the pseudo observations to the corresponding actual observations from the sample.

But we need more than the posterior distribution in order to formulate a credence about Mary. Let $\widetilde{X}_i \in \{0, 1\}$ be an additional outcome from group i (men or women) that has yet to be observed. This may be $X_{i(n+1)}$, $X_{i(n+2)}$ or $X_{i(n+m)}$ for any m between n and $N - n$ where uppercase N is the population size of group i. The inference about Mary is an inference about $\widetilde{X}_i$. The distribution of $\widetilde{X}_i$ given $x_{i1}, ...x_{in}$ is called the predictive distribution. For conditionally independent binary random variables, as we have here, this distribution can be derived from the distribution of $\widetilde{X}_i$ given θ_i and the posterior distribution of θ_i by averaging over the uncertain quantity θ_i. Following this approach, the probability that $\widetilde{X}_i = 1$ is given by,

$$\begin{aligned} p(\widetilde{X}_i = 1|x_{i1}, ..., x_{in}) &= \int p(\widetilde{X}_i = 1|\theta_i, x_{i1}, ..., x_{in})p(\theta_i|x_{i1}, ..., x_{in})d\theta_i \\ &= \mathrm{E}[\theta_i|x_{i1}, ..., x_{in}] \end{aligned}$$

[7] $f(x_i|\theta_i) = \binom{n}{x_i}\theta_i^{x_i}(1-\theta_i)^{n-x_i}$.

[8] A class P of prior distributions for θ is conjugate for a sampling distribution $f(x|\theta)$ if $p(\theta) \in P \rightarrow p(\theta|x) \in P$. The proof of this is straightforward and can be found in any introductory Bayesian statistics text, such as Gelman et al. (2013). Defining the class P requires a judgment call because if we define it broadly enough all distributions are conjugate to one another.

which is the posterior mean of θ_i, given by the sum of favorable pseudo and actual observations divided by the sum of all pseudo and actual observations. Therefore, the general schema for the posterior predictive distribution for binomial data is given by the following expression, which Huttegger (2017) calls the generalized rule of succession (we will see why below).

Generalized Rule of Succession.

$$p(\widetilde{X}_i = 1 | x_{i1}, ..., x_{in}) = \frac{\alpha + n\hat{\theta}_i}{\alpha + \beta + n}$$

and,

$$p(\widetilde{X}_i = 0 | x_{i1}, ..., x_{in}) = 1 - p(\widetilde{X}_i = 1 | x_{i1}, ..., x_{in})$$

Without further specifying the values of α and β there is little disagreement that this is the general form of the posterior predictive probability for $\widetilde{X}_i$. Huttegger (2017), following Zabell (2005), Carnap (1950) and Johnson (1924), shows that this form of the predictive probability follows from three very modest conditions. In particular, it follows if we assume that (a) the prior probabilities are regular (every finite sequence of outcomes has positive prior probability), (b) prior probabilities are exchangeable, and (c) the posterior predictive distribution of $\widetilde{X}_i = 1$ is a function of $\hat{\theta}_i$:

$$p(\widetilde{X}_i | x_{i1}, ..., x_{in}) = f(\hat{\theta}_i)$$

This last condition is known as Johnson's sufficientness postulate, after W.E. Johnson (Johnson, 1924; Zabell, 1982). The sufficientness postulate says that the predictive probability for $\widetilde{X}_i = 1$ (Mary is an admin) is a function of the number of female admins and the total sample size. The Generalized Rule of Succession extends naturally to multinomial data with any finite number of outcomes. We will stick to the binary case as in the Gender Bias Study for simplicity.

The real debate for our purposes is over the *extent* to which the sample determines our credence about Mary. In other words, it is over the particular values to be assigned to the parameters α and β. We can now restate the Frequency-Credence Principle. It requires, specifically, that we set $\alpha = 0$ and $\beta = 0$. In other words, $p(\widetilde{X}_i = 1 | x_{i1}, ..., x_{in}) = \hat{\theta}_i$. This requires not only that the inference be a function of the observed data but more specifically that it be given by the identity function $f(x) = x$. This is Reichenbach (1938, 1949)'s "Straight Rule".

From the accuracy-first perspective of epistemic rationality, the Frequency-Credence Principle, in its naive form, is epistemically irrational. This is because when $\alpha = \beta = 0$ the prior distribution for θ is given by $p(\theta) \propto 1/(\theta(1-\theta))$ which is incoherent because $\int_0^1 1/(\theta(1-\theta))d\theta \neq 1$. Indeed, it is worse than incoherent, because the integral does not converge, meaning that there is no normalizing constant that would turn this prior into a coherent probability function. This is not necessarily a criticism of Reichenbach, as he did not purport to give a Bayesian account of induction. Rather, it is to say that from our accuracy-first perspective the Straight Rule does not look very attractive because it presupposes prior credences which violate Joyce (1998, 2009)'s coherence norm.

However, consider Laplace's original Rule of Succession, which is a specification of the Generalized Rule of Succession with $\alpha = \beta = 1$. This is a better candidate because it assumes a uniform prior for θ_i which is coherent since $p(\theta_i) = 1$ integrates to 1 over the unit interval. If we use Laplace's Rule of Succession in Mary's case and identify our credence with the mean of the posterior probability distribution then the probability that Mary is a faculty member will be equal to $(15 + 1)/(50 + 2) = 0.31$. We suspect (and to be charitable, we will assume) that when it comes to predictive inference, what supporters of the Frequency-Credence Principle have in mind is Laplace's Rule of Succession. Since by definition we did not know anything about the distribution of different occupational roles in academia across gender, epistemic rationality requires that we start with a uniform prior before becoming aware of the study. Indeed, White (2010) defends the Frequency-Credence Principle together with Laplace's other well known maxim, the principle of indifference, which requires a uniform prior in the absence of information.

In the next section, we rely on a theory of epistemic risk articulated in Babic (2017) to ague that while the Generalized Rule of Succession is indeed a requirement of epistemic rationality, Laplace's specific Rule of Succession is not. Laplace's rule follows from a particular attitude to the costs of being mistaken, and this attitude is not a requirement of epistemic rationality. Different attitudes to the cost of error are permissible which in turn suggest different values for the parameters α and β in making a predictive inference about Mary.

5 Epistemic Risk

While there is no consensus on which precise functional form a measure of inaccuracy should take, there is wide agreement that a measure satisfying (1)-(3) is rationally required (e.g., Greaves and Wallace, 2006; Joyce, 2009; Huttegger, 2017). However, there are infinitely many functions satisfying this requirement. For example, adding a constant to squared Euclidean distance does not violate any of the constraints. Further, the constant can vary with the event. So a function of the form $(v - pr(h))^2 + k_v$ satisfies (1)-(3). These modifications are not normatively innocuous, however.

In accuracy based epistemology an agent's measure of inaccuracy can reflect her normative attitudes to the cost of approaching different types of error. In other words, it can reflect the agent's attitudes to epistemic risk. Accuracy, therefore, is not a purely alethic concern. Joyce (2015) and Levinstein (2017) highlight this point as well. This is a feature of the account rather than a bug. It is compatible with the accuracy framework for an agent to have pragmatic reasons for the particular way in which she values accuracy. For example, it is reasonable for a weather forecaster to care more about false negative mistakes when the hypothesis is 'there is a tornado nearby'.

5.1 Measuring Epistemic Risk

Suppose we have two decision makers, A and B. Neither A nor B has any prior information about the gender distribution across different occupational roles in academia, and neither has met Mary. Both A and B read about the Gender Bias Study one morning, and both encounter Mary on campus later that day. Their task is to assign a credence to the proposition $h :=$ 'Mary holds an administrative position'. Levinstein (2017) makes the following point:

> Plausibly, [epistemic] rationality permits a range of different attitudes toward epistemic risk. If [A and B] have different risk-profiles, then they can rationally maintain disagreement [with respect to h] (pg. 302).

However, Levinstein leaves open the following questions: What does it mean to have different attitudes to epistemic risk? How can we measure and compare different epistemic risk profiles? And how do these risk profiles affect A and B's credences about Mary. We will answer these questions.

Assume A is indifferent between approaching false positive and false negative mistakes regarding Mary's role. While she would like to reach an accurate judgment, she has no preference between falsely assuming that Mary is an admin (a false positive mistake with respect to h) or falsely assuming that Mary is a faculty member (a false negative mistake with respect to h). Meanwhile, B shares Blackstone's aversion to false positive mistakes. For her, falsely assuming that Mary is an admin is worse than falsely assuming that Mary is a faculty member in a world where faculty members are held in higher regard. Babic (2017) demonstrates that these attitudes to risk of error determine the agent's measure of accuracy up to an additive constant and, together with an attitude to how much risk they are willing to assume in an inference problem, they determine the agent's prior credence function. That is, they identify a unique $p \in \mathcal{P}$. As a result, they specify the particular form (or, at a minimum, set of permissible forms) for the Generalized Rule of Succession – they suggest the appropriate values for α and β.

Consider agent A. How can we understand her epistemic risk function? We know it is indifferent between changes in inaccuracy in the false positive/ false negative error direction. For simplicity, we may also suppose that it increases at an equal rate with the extremity of her credence. So we want a risk function whose derivative increases linearly in $pr(h)$. Finally, we know, from the equal regard to error assumption, that it must reach a minimum at $pr(h) = 0.5$. Indeed, with this credal value the agent would receive an equal payoff, in terms of inaccuracy, regardless of the actual outcome. We have sketched such a function in Figure 1. The x-axis represents the agent's credence in h whereas the y-axis represents her assessment of epistemic risk, $R(pr(h))$, associated with that credence in arbitrary units.

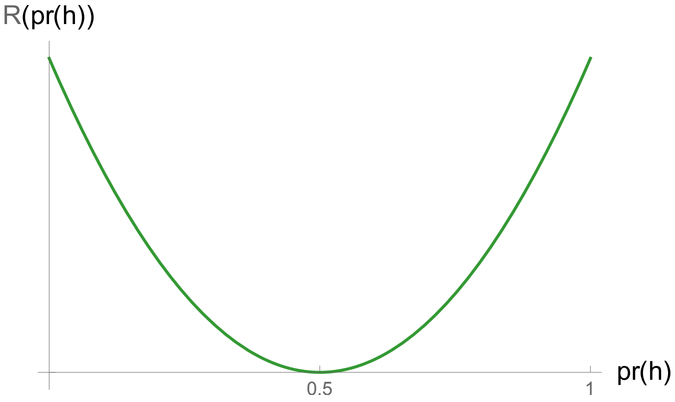

Figure 1: Symmetric epistemic risk function

Letting $x = pr(h)$, the function in Figure 1 is given by rescaling $-x(1-x)$. The scale is arbitrary. We know our function must reach its local minimum at $x = 0.5$. We have chosen to set that minimum to 0. Its derivative is $1 - 2x$ whose absolute value increases linearly in x, in a way that is symmetric around 0.5. Our point here is conservative. It is simply that this is a plausible epistemic risk function consistent with A's attitudes to the relative cost of moving in either error direction.

Consider agent B now. Since she is more sensitive to approaching false positive mistakes, her least risky point, where her inaccuracy if the proposition is false is exactly the same as her inaccuracy if the proposition is true, must be shifted to the left. That is, her safest point will be such that $pr(h)/pr(\overline{h}) < 1$. Suppose for the sake of illustration that this agent's epistemic risk function reaches a local minimum where $pr(h)/pr(\overline{h}) = 1/3$. Then the safest credence will be $pr(h) = .25$. We have sketched one such function in Figure 2.

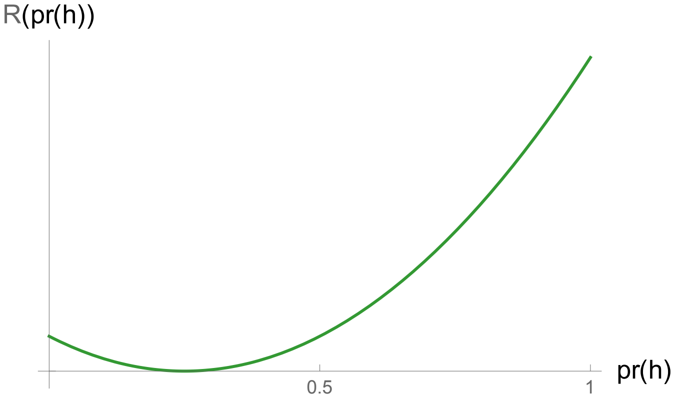

Figure 2: Non-symmetric epistemic risk function

Letting $x = pr(h)$ again, this function is obtained by rescaling $(x - 1/2)x - 1/2$. Its derivative is $2x - 1/2$, which increases linearly in x from its minimum value at .25. This means there is a lot more risk to assume, so to speak, as we approach higher credal values for h (and risk increasing false positive inaccuracy) than there is as we shift down to even lower credal values (and risk increasing false negative inaccuracy).

5.2 Risk and the Selection of Accuracy Measures

We hope the story so far is sufficiently compelling. We have stipulated certain facts about how sensitive A and B are to unit changes in inaccuracy in the direction of false positive and false negative mistakes. We have then used this information to plot a simple function that is consistent with A and B's attitudes to risk of error – their epistemic risk function. Babic (2017), following Savage (1971), demonstrates that if a risk function is (1) convex and (2) continuous on the unit interval, as the above functions are, then we can derive from it a unique scoring rule that satisfies the requirements (1)-(3) from above: it is truth-directed, continuous, and strictly proper. Thus, the associated scoring rule is at a minimum a permissible candidate for measuring inaccuracy from the perspective of Bayesian epistemic rationality. As a result, the normative judgments encoded in the risk function are therefore carried forward to the agent's scoring rule itself. So while scoring rules satisfying (1)-(3) may all be epistemically rational, they reflect different normative (moral, prudential) attitudes to the risk of graded error.

From the risk function in Figure 1, given by $-x(1-x)$, we can derive the following scoring rule (all derivations are provided in the Mathematical Appendix).

Risk-neutral scoring rule. Letting $pr(h) = x$,

$$s_1(x) = (1-x)^2 \qquad\qquad s_0(x) = x^2$$

This is the ordinary Brier score. These functions are depicted in Figure 3, where the agent's inaccuracy is given by the solid blue curve if h is true and by the dashed red curve if h is false. The x-axis represents the probability assigned to h and the y-axis represents epistemic disutility.

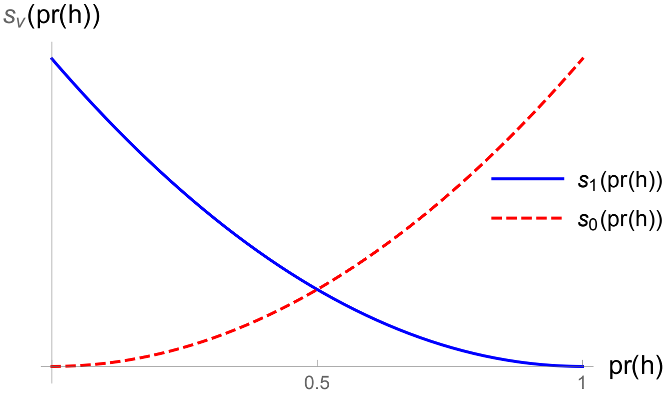

Figure 3: Risk neutral scoring rule derived from symmetric epistemic risk function

The least risky point, where the risk function $-x(1-x)$ reaches its minimum, is also the credal value where $s_1(x) = s_0(x)$. Further, there is a geometric relationship between the measure of epistemic risk (e.g. the function depicted in Figure 1) and its associated scoring rule (the functions depicted in Figure 3). The risk of any given credence corresponds to the area between s_1 and s_0, taken from the point of intersection to the vertical line extending from that credence. That is,

$$R(x) = \int_{x}^{x^*} |s_1(t) - s_0(t)| dt$$

where x^* satisfies $s_1(x) = s_0(x)$. This relationship is easier to grasp visually. It is depicted in Figure 4.

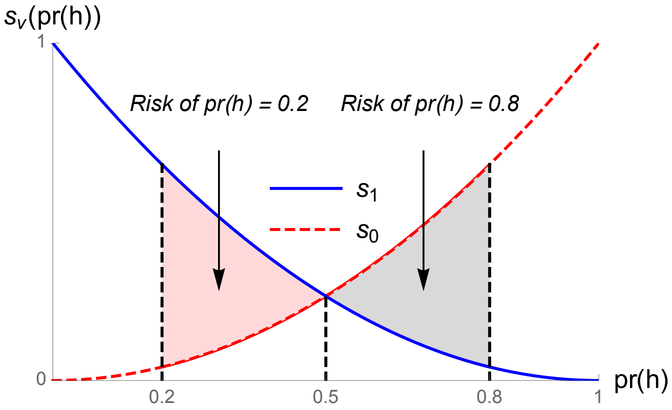

Figure 4: Relationship between scoring rule and epistemic risk

The shaded pink region (to the left of the intersection) corresponds to the risk of credence 0.2 in h and the shaded gray region (to the right) corresponds to the risk of credence 0.8 in h. Since this agent is indifferent between unit changes in inaccuracy in either direction, we can tell that these regions are equal in area. Moving 0.3 in either direction from 0.5 increases risk by the same amount. Therefore, we can see quite clearly that the agent's normative attitudes to the relative cost of error determine the shape of her scoring rule through this geometric relationship between the scoring rule and the epistemic risk function. By depicting this relationship in Figure 4, it is easy to see that this is a scoring rule under which approaching false positive inaccuracy is just as bad as approaching false negative inaccuracy.

Meanwhile, from the risk function in Figure 2, we can derive the following measure of inaccuracy.

Scoring rule with false-positive error bias. Letting $pr(h) = x$,

$$s_1(x) = (1-x)^2 \qquad\qquad s_0(x) = x^2 + 1/2$$

This is also a quadratic score. When inaccuracy increases in the false negative direction, it is identical to the Brier score, but when inaccuracy increases in the false positive direction there is an added penalty term. These functions are depicted in the left-panel of Figure 5, where the agent's inaccuracy is again given by the solid blue curve if h is true and by the dashed red curve if h is false. The least risky credal value, as we saw in Figure 2, is $pr(h) = .25$. The relationship between this scoring rule and its associated epistemic risk function is depicted in the right-panel of Figure 5.

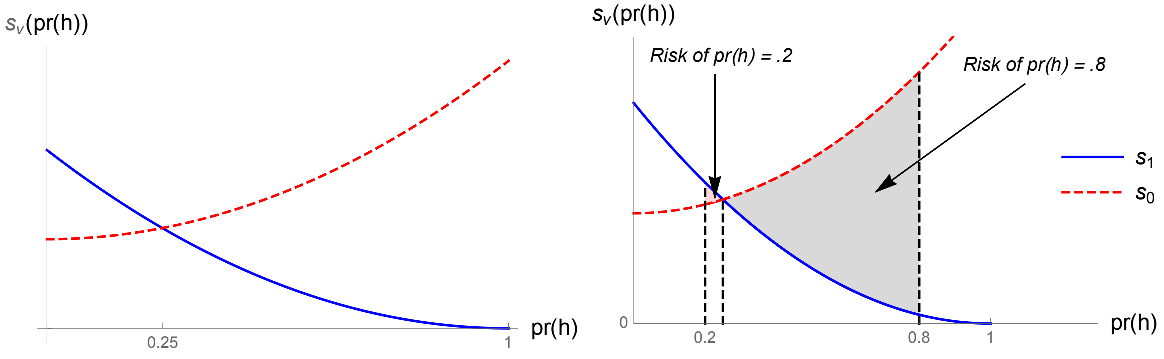

Figure 5: Non-symmetric scoring rule and its relationship to epistemic risk

For this agent, a 0.8 credence is much riskier than a 0.2 credence and moving from 0.5 to 0.8 is no longer the same as moving from 0.5 to 0.2. We can see that this is a scoring rule whose shape is such that approaching a false positive mistake is much worse than approaching a false negative mistake.[9] Each of the scoring rules considered in this section satisfies (1)-(3) – they are truth-directed, continuous, and strictly proper – because they are derived from a convex and continuous epistemic risk function.

5.3 Prior Probabilities and Aversion to Epistemic Risk

Even if we know the function that characterizes an agent's attitudes to epistemic risk, we cannot yet say what credences she ought to hold, either before or after making any observations. Since we have limited our attention to convex and continuous epistemic risk functions, which produce strictly proper scoring rules, whichever credences an agent selects before making any observations will consider themselves to be most accurate in expectation. Gibbard (2008) calls this the 'immodesty' of strictly proper scoring rules. Therefore, in order to make a further statement about what prior credences the agent should hold, before seeing evidence like the Gender Bias Study, we have to stipulate their attitude to how much epistemic risk they are willing to assume. In other words, we need to know their prior degree of epistemic risk aversion.

While the epistemic utility framework is compatible with many different attitudes to epistemic risk aversion, it has often been argued that an agent should determine their prior credence function by minimizing epistemic risk. For example, Pettigrew (2016) in effect argues for this position by suggesting that an agent should select as her prior credences the function $p \in \mathcal{P}$ that is *minimax dominant* as compared to all other $q \in \mathcal{P}$. That is, she should assign her prior credence function in a way that minimizes her worst case inaccuracy. In our simple setup, there are only two cases – either h is true or it is false.

Consider agent A first, whose measure of inaccuracy is given by the ordinary Brier score (Figure 3). Suppose A's credence for h is 0.6. Then if h is false she could have done

[9] Scoring rules are very flexible and may reflect many different attitudes to epistemic risk. We use the easiest case here for illustration, where the loss is simply shifted up in the false positive error direction.

better by going down to 0.5. However, if she reduces her credence further, then if h is true she can do better by moving up to 0.5. By reasoning in this way she will arrive at an equilibrium credence of 0.5 where her payoff is the same regardless of the outcome. This credence is minimax dominant because she can no longer improve her epistemic situation in one outcome without making it worse in the other. More generally, $pr(h) = 0.5$ is minimax dominant for the Brier score *because* $s_1(0.5) = s_0(0.5)$. But this is not in general true of the uniform credence $pr(h) = 0.5$. In other words, $pr(h) = 0.5$ is *not* minimax dominant *because* it is uniform. This is an important distinction. Sometimes uniformity and epistemic risk aversion go together, as in the case of ordinary squared distance, but often they do not, as in the penalized case of squared distance.

In any case, while there is little reason to believe that an agent is rationally required to be maximally epistemically risk averse in the absence of evidence it is at least rationally permissible to adopt this attitude when identifying a prior. Our argument does not depend on any particular degree of epistemic risk aversion being the right one. We simply need to stipulate our agents share some attitude to epistemic risk aversion – any attitude – in order to explore how their beliefs change after they receive evidence. For simplicity, we will assume our agents are epistemic risk-minimizers.

Consider now our agent B, whose measure of inaccuracy is given by the penalized score $(v - pr(h))^2 + k_v$ where $k_v = 0$ when $v = 1$ and $k_v = 1/2$ when $v = 0$ (Figure 5). For this agent, the credence that satisfies the equation $s_1(x) = s_0(x)$ is $pr(h) = .25$. Therefore, for B this is the minimax dominant credence or, equivalently, the prior credence function she would select if she were an epistemic risk minimizer. It is not uniform.

To sum up: we have two agents, A and B. A's measure of inaccuracy (Figure 3) is derived from the epistemic risk function depicted in Figure 1, while B's measure of inaccuracy (Figure 5) is derived from the epistemic risk function given in Figure 2. A and B share the same attitude to epistemic risk – they are maximally risk averse. This implies that in the absence of evidence they select the credence that minimizes their worst case inaccuracy. For person A this credence function is given by $(.5, .5)$ in h and its negation. For person B this credence function is given by $(.25, .75)$ for h and its negation.

6 Avoiding Normative Conflicts

We have seen how normative considerations – assessments about the relative severity of approaching different types of error – determine the shape of an agent's measure of inaccuracy. We now apply this insight to the Gender Bias Study to see how an agent can avoid normative conflicts.

Consider A, first. Her prior credence, before observing any evidence, that Mary is an admin is 0.5. This follows from the assumption that she is an epistemic risk minimizer whose epistemic risk function treats all errors equally. These attitudes to epistemic risk require a prior probability for the unknown continuous rate parameter θ_2 (the population frequency of female administrators) given by $p(\theta) \propto 1$ (depicted as the solid green curve in the left-panel of Figure 6). If $p(\theta) \propto 1$ then A's prior beliefs correspond to a beta distribution with $\alpha = 1$ and $\beta = 1$. Moreover, in the absence of evidence, the Generalized Rule of Succession requires that the prior predictive probability regarding Mary's occupation is given by $\alpha/(\alpha+\beta)$. With a $\text{Beta}(1, 1)$ distribution this is 0.5.

But why not $\text{Beta}(2, 2)$ or $\text{Beta}(5, 5)$? These alternatives are not permitted from A's perspective because we have supposed that A is the kind of agent who treats every type of error equally in the absence of information. If A were to assign any non-uniform prior distribution to θ_2, this would suggest that A is more sensitive to being mistaken with respect to certain possible true values of θ_2 than others. For example, even when $\alpha = \beta$ any prior other than $\text{Beta}(1, 1)$ implies that our agent is extra sensitive to being mistaken when θ_2 is near 0.5.

So if A believes her probabilities about the unknown proposition h should be uniform, because she is an epistemic risk-minimizer with a symmetric epistemic risk function, then she likewise believes that her probabilities about the unknown quantity θ_2 should be uniform, due to the same considerations. The only case where both distributions are uniform is where the prior for θ_2 follows a $\text{Beta}(1, 1)$ distribution. As a result, A's posterior for θ_2 after becoming aware of the Gender Bias Study must be $\text{Beta}(1 + 15, 1 + 35)$ (the dashed red curve in the left panel of Figure 6). Therefore, A would apply Laplace's Rule of Succession in making a predictive inference about Mary given by, $pr(h) = 36/52 = 0.69$. A is quite confident that Mary is an administrator but slightly less so than she would have been had she applied Reichenbach's Straight Rule. Given these assumptions, A appears to be under Normative Conflict.

But we are now in a position to understand why these strong assumptions are not requirements of epistemic rationality on the accuracy framework. Consider agent B. We know that her safest prior credence function, before seeing any evidence, in the proposition that Mary is an administrator is .25.[10] Using the same link, as above, between $pr(h)$ and $p(\theta_2)$, this implies that for B any prior for θ_2 which satisfies $\text{E}[\theta_2] = .25$ is a permissible prior to adopt. We have not said anything else about B's attitudes to epistemic risk that would further constrain her set of permissible priors for θ_2. We know that the posterior is given by $\text{Beta}(\alpha + 35, \beta + 15)$. Suppose that for B, $\alpha = 10$ and $\beta = 30$ (the solid green curve in the right-panel of Figure 6). This is consistent with B's attitudes to epistemic risk because in this case $\text{E}[\theta_2] = 10/40 = .25$. This leads to a $\text{Beta}(45, 45)$ posterior (the dashed red curve in the right panel of Figure 6). B now has to formulate a credence as to whether Mary is an administrator. While she is not going to use Laplace's Rule of Succession, she will indeed apply Huttegger (2017)'s Generalized Rule of Succession. Using this principle, her credence that Mary is an administrator is $pr(h) = 45/90 = 0.5$. B has avoided Normative Conflict.

[10] We would like to anticipate a potential objection here. It may seem unusual to suppose that A and B have precise credences about Mary before meeting her. This is not essential to the account. Rather, we want to evaluate A and B's posterior credences about Mary, after they meet her, from a normative perspective. Our point here is that B's attitude to risk reflects the set of prior credences she would have considered appropriate to hold before meeting Mary.

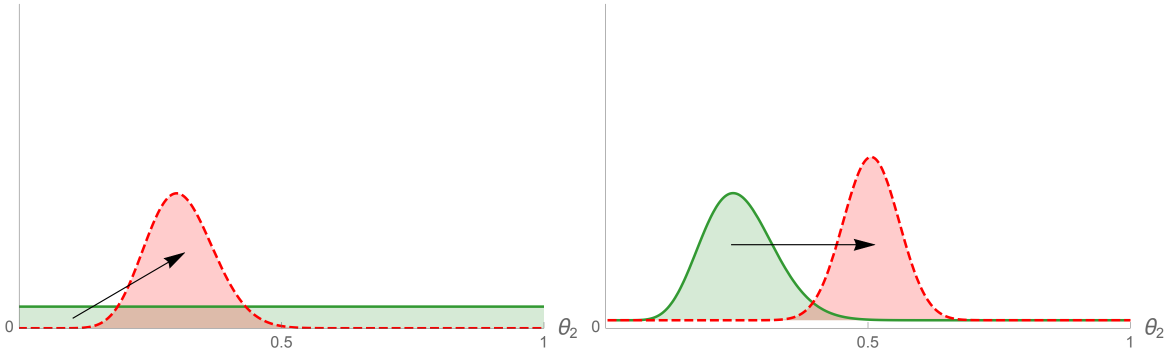

Figure 6: Priors (solid green) and posteriors (dashed red) for θ_2. A (left-panel) and B (right-panel). The black arrows represent Bayesian updating on the Gender Bias Study.

Neither A nor B have violated any principles of epistemic rationality. Both A and B started with a coherent prior, which they updated by Bayesian conditionalization. Furthermore, both A and B applied Huttegger's Generalized Rule of Succession in order to make a predictive inference about Mary and formulate a credence as to whether she is a faculty member. A is quite sure that Mary is not a faculty member whereas B is equally undecided.

Moreover, at no point did their credences change in the absence of evidence. We do not endorse considerations of epistemic risk as a decision-rule for what to believe at any point in time. Rather, an agent's posterior credences reflect their prior attitude to epistemic risk. In this case, if you, like many people, are hesitant to assume that Mary is not a faculty member on the basis of the Gender Bias Study our conclusion is that you are permitted to adopt a middling credence. This credence is consistent with having a coherent prior, updating by Bayesian conditionalization, and applying the Generalized Rule of Succession. What this credence says about your attitudes to epistemic risk, though, is that you are extra sensitive to falsely assuming that Mary is not a faculty member.

We have not filled in the details of why you are extra sensitive to falsely assuming that Mary is not a faculty member. It may be that having read Basu and Schroeder (2018), you want to avoid wronging Mary by your beliefs about her. It is also possible that you are worried about acting on the basis of your beliefs and perhaps saying something offensive to Mary. It is likewise possible that your concerns are purely prudential. There is nothing in the accuracy framework that rules out these normative considerations affecting your attitudes to epistemic risk and in turn your scoring rule and inferential behavior.

Therefore, the Sad Conclusion that Gendler (2011) describes only applies to agents who apply a parity principle to different types of error and Laplace's Rule of Succession in making a predictive inference. But this is a substantive normative position – it is not a requirement of epistemic rationality. When we risk falsely accusing someone on the basis of their gender, ethnicity, or skin color, it is morally appropriate, and epistemically permissible, to reject such a parity principle and to make a predictive inference using a version of the Generalized Rule of Succession that is not Laplacean.

The argument presented here should be interpreted with care. It is limited to what we take to be ordinary cases giving rise to Normative Conflict. We have used an example that is representative of prevailing empirical work in leading psychology journals. Still, this leaves open the possibility that a massive study involving thousands of individuals reinforces a pernicious stereotype thereby giving rise to Normative Conflict. Making a prediction on the basis of such a result would not be altogether different from making a Direct Inference. In such cases, the Frequency-Credence Principle fares better and the relevant question would be whether the individual about whom we are formulating a credence has been randomly selected for presentation. Fortunately, the random selection condition is often not satisfied in cases that might give rise to normative conflicts. The way an individual is presented to us is often informative, which means that our prior probabilities about all such individuals are not exchangeable. For example, suppose you run into Mary before a colloquium, she looks quite professorial and is carrying her laptop and some notes. This is information that a prediction about Mary's occupation should take into account. In this case, even if the Gender Bias Study had a much larger sample size that overpowered our prior, we would now have to update on the new information about Mary. Therefore, even if the sample size is large and the research conducted is beyond assail, if other information becomes available, as it often does, normative conflicts need not arise. Finally, we have assumed the study was competently performed. This assumption is not necessary to the argument – rather, we think it is obvious that normative conflicts need not occur where the frequency arises from poorly conducted research. Indeed, this is often the case with empirical work giving rise to normative conflicts – experiments suggesting racial or ethnic disparities due to biased collection methods or other design flaws.

7 Concluding Remarks

We have argued that an agent who is worried about wronging someone like Mary by adopting a credence about her, after becoming aware of evidence like the Gender Bias Study, is ordinarily epistemically permitted to adopt a prior probability that reflects this worry. For such an agent, there will be many coherent prior distributions which are consistent with these attitudes to epistemic risk. These prior distributions will be such that after updating by conditionalization and applying the Generalized Rule of Succession, her credence about someone like Mary will not be wrongful.

If the argument developed in this paper still seems objectionable, one way to interpret this essay is as an invitation to explain what exactly someone like person B has done wrong from the perspective of epistemic rationality. For example, one approach would be to argue that epistemic rationality requires an agent to be indifferent between approaching different types of mistakes. Pettigrew (2016) suggests this approach while Levinstein (2017) explicitly rejects it. It is hard to see how we could derive such a substantive requirement about attitudes to risk of error from epistemic considerations alone. Another approach would be to put an upper bound on the weight of prior beliefs in predictive inference. For example $\alpha + \beta \leq 5$. In this case, an agent would be free to adjust her prior credences in a way that reflects her attitudes to epistemic risk given a fixed "budget" of weight that can be distributed to her priors. But the determination of her budget seems ad hoc and insufficiently motivated from an epistemic perspective. Why 5 instead of 15 or 150? Alternatively, we could impose an upper bound on the weight of priors in predictive inference that is sensitive to sample size. For example, $\alpha + \beta \leq \sqrt{n}$. This is somewhat more plausible as it resembles in some respects an approach to model assessment called robustness analysis in Bayesian statistics. The idea of a robustness analysis is to examine how one's inferences change across different plausible specifications of the prior and sampling distributions. But even in a robustness analysis it is generally a bad idea to take any given rule of thumb and treat it as an inferential requirement.

Mathematical Appendix

Deriving risk-neutral scoring rule.

Let $pr(h) = x$, where h and its negation form a partition of Ω. Let $p = (x, 1-x)$ be a probability distribution on Ω. We are given that $R(x) = -x(1-x)$. Further let $H(x) = E_p[s_v(x)]$, the expected inaccuracy of p calculated using p itself.

Savage (1971) shows that if the measure of inaccuracy s is strictly proper, we can express it in terms of the expected inaccuracy function H as follows,

$$s_1(x) = H(x) + (1-x)H'(x) \qquad\qquad s_0(x) = H(x) - xH'(x)$$

Babic (2017) shows that $R(x) = k - H(x)$ where $k = \max_x H(x)$. Recall that in the main text we set the scale to 0 for convenience. The scaling constant that sets the scale to 0 is equal to $\max_x H(x)$. Therefore, we can derive the scoring rule as follows:

$$\begin{aligned}
s_1(x) &= H(x) + (1-x)H'(x) \\
&= k - R(x) + (1-x)\frac{d}{dx}[k - R(x)] \\
&= k - R(x) + (1-x)(1-2x) \\
&= k - (k - x(1-x)) + (1-x)(1-2x) \\
&= x(1-x) + (1-x)(1-2x) \\
&= (1-x)^2
\end{aligned}$$

$$\begin{aligned}
s_0(x) &= H(x) - xH'(x) \\
&= k - R(x) - x\frac{d}{dx}[k - R(x)] \\
&= x(1-x) - x(1-2x) \\
&= x^2
\end{aligned}$$

Since this is the ordinary Brier score, it is well-known to be truth-directed, continuous and strictly proper.

Deriving scoring rule with false-positive error bias.

In this case, $R(x) = (x - 1/2)x - 1/2$. We can again derive the scoring rule as follows:

$$\begin{aligned} s_1(x) &= H(x) + (1-x)H'(x) & s_0(x) &= H(x) - xH'(x) \\ &= k - R(x) + (1-x)\frac{d}{dx}[k - R(x)] & &= k - R(x) - x\frac{d}{dx}[k - R(x)] \\ &= -((x-1/2)x - 1/2) + (1-x)(1/2 - 2x) & &= -((x-1/2)x - 1/2) - x(1/2 - 2x) \\ &= (1-x)^2 & &= x^2 + 1/2 \end{aligned}$$

Since this score is obtained by rescaling the Brier score under s_0 it remains truth-directed and continuous. We can also confirm it is strictly proper.

Let $q = (b, 1-b)$ be the agent's credence function on Ω. Let $I(b,x) = b(1-x)^2 + (1-b)(x^2 + 1/2)$ be the expected inaccuracy of distribution p calculated using the credences q. The first and second derivatives if $I(b,x)$ are:

$$\frac{\partial}{\partial x} I(b,x) = 2(x-b) \qquad\qquad \frac{\partial^2}{\partial x^2} I(b,x) = 2$$

Therefore, $I(b,x)$ is uniquely minimized when $b = x$.

References

Babic, B. (2017). *Foundations of Epistemic Risk*. Ph. D. thesis, University of Michigan, Ann Arbor, available at https://deepblue.lib.umich.edu/handle/2027.42/140922.

Basu, R. (2018). The Specter of Normative Conflict. In E. Beeghly and A. Madva (Eds.), *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*. London: Routledge.

Basu, R. and M. Schroeder (2018). Can Beliefs Wrong? In *Philosophical Topics (Special Issue)*. Forthcoming.

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

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

De Finetti, B. (1974). *Theory of Probability*, Volume 1. New York: John Wiley and Sons.

Easwaran, K. (2013). Expected Accuracy Supports Conditionalization – and Conglomerability and Reflection. *Philosophy of Science 80*(1), 119–142.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis* (3rd ed.). New York: CRC Press (Taylor & Francis).

Gendler, T. S. (2011). On the Epistemic Costs of Implicit Bias. *Philosophical Studies 156*(1), 33–63.

Gibbard, A. (2008). Rational Credence and the Value of Truth. In *Oxford Studies in Epistemology*, Volume 2. Oxford: Oxford University Press.

Greaves, H. and D. Wallace (2006). Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind 115*(459), 607–632.

Holmes, C. B., J. M. Marszalek, C. Barber, and J. Kohlhart (2011). Sample Size in Psychological Research Over the Past 30 Years. *Perceptual and Motor Skills 112*(2), 331–348.

Huttegger, S. M. (2013). In Defense of Reflection. *Philosophy of Science 80*(3), 413–433.

Huttegger, S. M. (2014). Learning Experiences and the Value of Knowledge. *Philosophical Studies 171*(2), 279–288.

Huttegger, S. M. (2017). *The Probabilistic Foundations of Rational Learning.* Cambridge: Cambridge University Press.

Johnson, W. (1924). *Logic, Part III. The Logical Foundation of Science.* Cambridge: Cambridge University Press.

Joyce, J. M. (1998). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science 65*, 575–603.

Joyce, J. M. (2009). Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In F. Huber and C. Shmidt-Petri (Eds.), *Degrees of Belief*, pp. 263–300. Springer.

Joyce, J. M. (2015). Prior Probabilities as Expressions of Epistemic Values. Draft.

Kyburg, H. E. (1974). *The Logical Foundations of Statistical Interference.* Dordrecht: Reidel.

Leitgeb, H. and R. Pettigrew (2010). An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science 77*(2), 236–272.

Levi, I. (1977). Direct Inference. *The Journal of Philosophy 74*(1), 5–29.

Levinstein, B. (2017). Permissive Rationality and Sensitivity. *Philosophy and Phenomenological Research 94*(2), 342–370.

Mogensen, A. (2018). Racial Profiling and Cumulative Justice. *Philosophy and Phenomenological Research* (Early View).

Pettigrew, R. (2016). Accuracy, Risk, and the Principle of Indifference. *Philosophy and Phenomenological Research 92*(1), 35–59.

Railton, P. (1994). Truth, Reason, and the Regulation of Belief. *Philosophical Issues 5*, 71–93.

Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge.* Chicago: University of Chicago Press.

Reichenbach, H. (1949). *A Theory of Probability.* Berkeley: University of California Press. Original German edition, 1935.

Savage, L. J. (1971). Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association 66*(336), pp. 783–801.

Tetlock, P. E., O. V. Kristel, S. Elson, M. C. Green, and J. S. Lerner (2000). The Psychology of the Unthinkable: Taboo Trade-Offs, Forbidden Base Rates, and Heretical Counterfactuals. *Journal of Personality and Social Psychology 78*(5), 853–870.

van Fraassen, B. C. (1995). Belief and the Problem of Ulysses and the Sirens. *Philosophical Studies 77*(1), 7–37.

Weatherson, B. (2014). Running Risks Morally. *Philosophical Studies 167*(1), 141–163.

Wedgwood, R. (2002). The Aim of Belief. *Philosophical Perspectives 16*(s16), 267–297.

White, R. (2010). Evidential symmetry and mushy credence. In T. S. Gendler and J. Hawthorne (Eds.), *Oxford Studies in Epistemology*, Volume 3, Chapter 7, pp. 161–186. Oxford: Oxford University Press.

Zabell, S. L. (1982). W.E. Johnson's Sufficientness Postulate. *The Annals of Statistics 10*(4), 1090–1099.

Zabell, S. L. (2005). *Symmetry and its Discontents: Essays on the History of Inductive Probability.* Cambridge: Cambridge University Press.