Moral Obligation and Epistemic Risk

Boris Babic (INSEAD) and Zoë Johnson King (NYU)

Forthcoming in **Oxford Studies in Normative Ethics**, Vol. 10 (2019)

1. The Good News

This paper brings good news.

To explain the good news, we first need to explain a bit of purported bad news. In 2011, Tamar Gendler argued 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 (p.57).

Gendler calls this "the sad conclusion" (Gendler 2011).

It's easy to understand Gendler's concern. Many personal traits — such as prior criminal convictions, low socioeconomic status, and certain medical conditions — are subject to profound social stigma. And we often become aware of information about the base rates of these socially disvalued traits within a certain (gender, racial, ethnic, etc.) group. It is natural to think that epistemic rationality then requires us to revise our estimate of the probability that members of the group whom we encounter possess the trait, in line with our information about the base rate of the trait within the group. Many formal epistemologists would say, and many other theorists assume, something stronger: that epistemic rationality requires us to match our estimates of the probability that individual group members possess the trait to the base rates that we have so far come across.

Here is an example:

> **Gender Bias Study.** One morning, in your usual newspaper (which you take to be reliable), you read a report about a study on gender discrepancies in academic employment. The study surveyed 500 men and 500 women employed by higher education institutions similar to your local university. 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: 70% were faculty, and 30% were administrative assistants. You are intrigued by this, since, before reading about the study, you had no information about the distribution of employment roles across gender lines within higher education. But it appears that the study was conducted competently; the number of participants was stipulated

> in advance, all identified covariates were tracked, no observations were excluded, and so on. Later that day you meet Mary, a new neighbor. Mary tells you that she is about to start working at your local university. But she does not tell you in what capacity she is to be employed.

What should be your estimate of the probability that Mary is a faculty member?

Someone who accepts Gendler's sad conclusion would say that this case involves a tragic conflict between the requirements of epistemic rationality and those of morality. They would say that epistemic rationality requires you to hold that Mary is 70% likely to be an administrative assistant and 30% likely to be faculty. After all, these are the probabilities suggested by the study you just learned about. And there is no reason to doubt its reliability. Moreover, you have no other information bearing on the probability that a woman employed at your local university is a faculty member besides the information that was reported in the study. Nonetheless, these assumptions about Mary seem morally problematic. They reinforce cultural associations between being a woman and being in low-paid, low-status jobs, fueling a hostile environment for working women. They also might lead you to act in ways that wrong Mary (for example, by talking down to her). And some philosophers think that we can wrong people just by believing negative things about them (see e.g. Basu and Schroeder 2018), in which case presumably we can wrong people to *degrees* by assigning degrees of belief to negative claims about them, which presumably holds of your expectation that Mary is an administrative assistant rather than a faculty member. Thus we arrive at a conflict: it is impossible to meet both the requirements of epistemic rationality and those of morality in your interaction with Mary.

This example concerns gender and employment. Gendler's example concerns race and crime rates. But the structure of the problem is the same. Indeed, structurally similar problems arise for all manner of social groups — based on religion, sexual orientation, disability, and so on — and all sorts of socially disvalued traits. The sort of case we are interested in here is one in which someone must estimate the probability that a particular person possesses a socially disvalued trait, based on information about the prevalence of that trait within a sample of a group to which the person belongs. In statistics, this estimation is called a *predictive inference.* When predictive inference requires someone to estimate that an individual they encounter is likely to bear a socially disvalued trait, we will here call it a *pernicious predictive inference.*

Here's the good news: Gendler's sad conclusion is false. It is false that there is a tragic conflict between the requirements of epistemic rationality and those of morality in cases like **Gender Bias Study**. This is because it is false that epistemic rationality requires you to assume that Mary is 70% likely to be an administrative assistant and 30% likely to be faculty. Indeed, epistemic rationality does not even require you to assume that Mary is more likely to be an administrative assistant than to be faculty. More generally, pernicious predictive inferences are not epistemically required. On the contrary, *no* particular attitude is required by epistemic rationality in these cases. And the sort of epistemic behavior that seems morally good — that is, a reluctance to draw pernicious predictive inferences about people based on information about the base rates of socially disvalued traits in groups to which they belong — is epistemically permitted. So there are all-things-considered normatively attractive options in cases like these;

options on which morality smiles, and at which rationality shrugs. Moreover, we can secure this result without adopting a conception of epistemic rationality that is unfriendly toward Bayesian statistical inference. On the contrary, we can show that pernicious predictive inferences are ordinarily not epistemically required even from a classical Bayesian perspective on the requirements of epistemic rationality, provided that we are clear about how such an approach measures and evaluates accuracy. That is our task in this paper.

This bit of good news is quite different from some other bits of good news that philosophers have recently reported on this topic. We are far from being the first to discuss pernicious predictive inferences; they have been subject to extensive discussion in philosophy of law for many years, and have recently become a hot topic for philosophers interested in the relationship between moral and epistemic norms. Philosophers in this literature have already shown that there are several ways in which pernicious predictive inferences may be epistemically problematic, as well as morally problematic. For example, they may misrepresent the generality of a statistic (Munton *ms*); they may lead to unhedged beliefs about an individual's possessing a trait when only a probabilistically hedged belief is warranted by the evidence (Gardiner *ms*); they may lead people to form beliefs without ruling out salient alternatives, where the standard for what it takes to rule out salient alternatives is raised by the morally charged nature of the case (Moss 2016, Chapter 10); or they may lead people to form beliefs based on insufficient evidence, where the threshold for sufficiency of evidence is raised by the morally charged nature of the case (Basu 2019). We do not challenge any of these arguments here. On the contrary, we take these philosophers to have identified a promising range of ways to find epistemic fault with pernicious predictive inferences.

What is striking about these extant arguments is that all they all focus on the negative: they take the sort of predictive inference that seems morally bad, and avoid the sad conclusion by arguing that it is also epistemically prohibited. By contrast, we focus on the positive: we take the sort of predictive inference that seems morally good, and vindicate its epistemic status.

This is important in part because it provides a way to avoid the sad conclusion that has been overlooked until now. But the sad conclusion is just as false if morally good predictive behavior is epistemically permitted as it is if morally prohibited predictive behavior is epistemically prohibited. The sad conclusion is that there is a tragic conflict between the requirements of epistemic rationality and those of morality in cases like **Gender Bias Study**. This is false if it turns out that rationality prohibits what morality prohibits, but equally false if it turns out that rationality permits what morality requires. In either case, there is no conflict.

Vindicating the epistemic status of morally praiseworthy inference is important for another reason, too. In our current social context, the ideas that morally praiseworthy predictive behavior is epistemically flawed is sometimes used as a smokescreen behind which prejudicial attitudes hide. Prejudiced people sometimes try to give the impression that cold, calculating epistemic rationality is on their side, and that reluctance to infer pernicious things about individuals based on statistical information about groups to which they belong is a sign of weakness or a denial of the "hard facts". This suggestion is worth challenging in and of itself, irrespective of the merits of other philosophical projects developing epistemic criticisms of pernicious predictive

inferences. We challenge the suggestion here. We hold that epistemic rationality permits morally praiseworthy predictive behavior. Moreover, we hold that this is a straightforward consequence of an intellectually honest Bayesian approach to epistemic rationality — i.e., one that is transparent about the underlying evaluative commitments behind competing ways of measuring accuracy. We take this dialectical point to be a particularly important contribution of our paper.

2. Predictive Inference

The conception of epistemic rationality that we employ in this paper is a simple *accuracy-based* account. This account's central assumptions are that the attitude of belief comes in degrees, which we can model using real numbers from 0 to 1 called *credences*, and that higher credences are more accurate than lower ones if the proposition in question is true, while lower credences are more accurate if it is false. On this approach, epistemic rationality requires that we form and revise credences with the aim of minimizing inaccuracy.

We measure inaccuracy using a function called a *scoring rule*, whose inputs are credences and outputs are "scores" of the credences' inaccuracy. There is consensus in the literature that good scoring rules have three properties: they are *monotonic*, *continuous*, and *strictly proper*. A monotonic scoring rule assigns credences a progressively higher inaccuracy score as they get progressively further from the truth. A continuous scoring rule is one whose scores increase smoothly as credences get further from the truth, without any small changes to someone's credence resulting in big jumps in their inaccuracy score. And a strictly proper scoring rule is one that, when used by an agent to compare her credences to others, always assesses her own current credences as uniquely best in expected accuracy. Beyond these three core constraints, there is little consensus on optimal scoring rules, and we will not assume anything further. But infinitely many scoring rules meet these three core constraints, and thus are acceptable from the point of view of accuracy-based epistemology. When epistemologists need an example of a kosher scoring rule, they often use squared Euclidean distance — a measure of inaccuracy proposed by the meteorologist Glenn Brier for assessing weather forecasts, and now known as the "Brier score". However, the Brier score is only one instance of a more general family of quadratic scoring rules. The Brier score, multiplied by any constant, would be likewise continuous, monotonic, and strictly proper.

Accuracy-based epistemologists place two requirements on credences. The first is that, at any point in time, an agent's credences must obey the *probability axioms*. Consider a *partition* of logical space: a set of mutually exclusive and jointly exhaustive possibilities. The probability axioms state that an agent's credence in the disjunction of any elements in a partition should be equal to the sum of her credences in each disjunct — for example, Zoë's credence that Boris is either at the gym or at the department should be the sum of her credence that Boris is at the gym and her credence that Boris is at the department — and moreover that her credences in all the elements of a partition should together sum to 1. The probability axioms also state that each of an agent's credences should be greater than or equal to 0. We call agents whose credences obey the probability axioms *probabilistically coherent.* Joyce (2009) vindicates the accuracy-based epistemologist's preference for probabilistically coherent credences by proving that, for any

incoherent credence function, there is a coherent one that is guaranteed to have a better accuracy score no matter which proposition turns out to be true, according to any scoring rule that is monotonic, continuous, and strictly proper. For epistemic agents who aim to minimize inaccuracy, then, coherent credences are always a better bet than incoherent ones.

The second central requirement within accuracy-based epistemology is that agents update their credences by *Bayesian conditionalization.* For example, suppose you learn that someone rolled a regular six-sided die. Your credence that it landed on 6 is 1/6. But if you were to learn that it landed on an even number, your credence that it landed on 6 would increase to 1/3. This 1/3 credence that the die landed on a 6, given that it landed on an even number, is called a *conditional credence*. To update by Bayesian conditionalization is to respond to new information (e.g. that the die landed on an even number) by shifting your prior credence in a proposition (e.g. your 1/6 credence that the die landed on 6) to your prior conditional credence in the proposition, conditional on the information that you just learned (e.g. your 1/3 credence that the die landed on 6, given that it landed on an even number). Just as Joyce has vindicated probabilistic coherence, Greaves and Wallace (2006) have shown that updating by conditionalization minimizes the prior expected inaccuracy of an agent's posterior credences, according to any scoring rule that is monotonic, continuous, and strictly proper. This means that the best strategy for an agent who aims to minimize inaccuracy when responding to new evidence is to update by Bayesian conditionalization.

To see how this approach understands predictive inference, consider the formal epistemologist's favorite illustrative example: a coin of an unknown bias. Suppose that Zoë is about to toss a coin, and Boris must guess whether it will land Heads. Boris knows that the probability of the coin landing Heads depends on its weight distribution and associated center of gravity; it might be a fair coin, with its weight distributed such that it lands heads 50% of the time, or it might be biased such that it lands heads 80% of the time, or 20%, etc. Boris doesn't know the coin's weight distribution. But he can assign credences to a range of hypotheses about the value of this *unknown parameter*. The possible values range from 0 (0% Heads) to 1 (100% heads), so Boris can *distribute* his credence over this range of hypotheses. For example, here are two possible distributions:

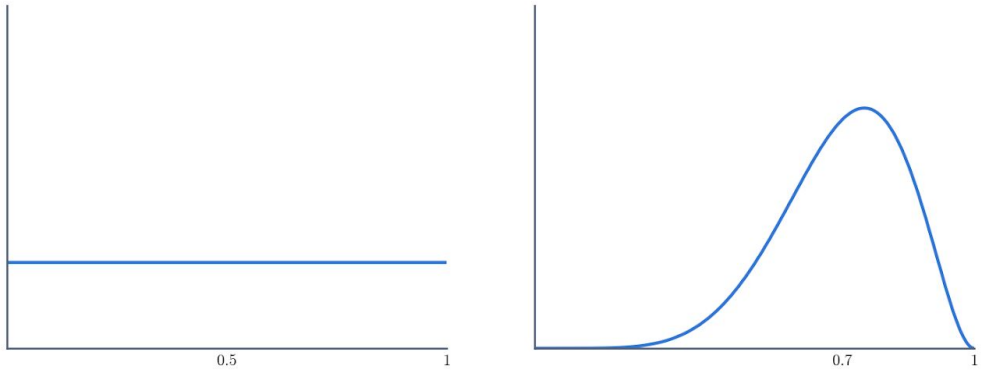

Fig. 1: Two probability distributions for coin's unknown bias

The distribution on the left is *uniform:* it assigns the same probability to every hypothesis about the coin's bias. The one on the right is *non-uniform*: it assigns some hypotheses a higher probability than others. But both distributions meet all the requirements of epistemic rationality on an accuracy-based approach.

When Boris makes a prediction about the probability that Zoë's coin toss will land Heads, he takes the *mean* of his current credal distribution over hypotheses about the underlying bias of the coin. For the distribution in the left panel in Fig. 1, the mean is 0.5, so Boris will guess that the coin is equally as likely to land Heads as it is to land Tails. For the distribution in the right panel, the mean is 0.7, so Boris will guess that the coin is about 70% likely to land Heads and 30% likely to land Tails.

Suppose that Boris sees Zoë toss the coin a few times, and sees whether it lands Heads or Tails each time. This gives him some data to use in his estimate of the coin's bias. Now, there's a heuristic we can employ to explain how Boris' credal distribution over hypotheses about the bias should change if he updates by Bayesian conditionalization: each probability distribution corresponds to a pair of numbers of Heads tosses and Tails tosses, as if Boris had set his initial credal distribution based on having observed precisely these numbers of tosses. For example, the uniform distribution in the left panel of Fig. 1 corresponds to one Heads toss and one Tails toss, while the distribution in the right panel corresponds to 7 Heads tosses and 3 Tails tosses. If he updates by Bayesian conditionalization, Boris' posterior credal distribution will correspond to the pair of numbers that would result from adding the number of Heads tosses that he actually observes to these "pseudo" initial Heads tosses, and then adding the number of Tails tosses that he actually observes to the "pseudo" initial Tails tosses. For example, if Boris were to start with the distribution in the right panel of Fig. 1, and then to observe 5 further Heads tosses and 3 Tails tosses, then his posterior distribution would correspond to 7+5 Heads and 3+3 Tails tosses.

The following diagrams show how the two distributions would shift after observing 5 Heads and 3 Tails:

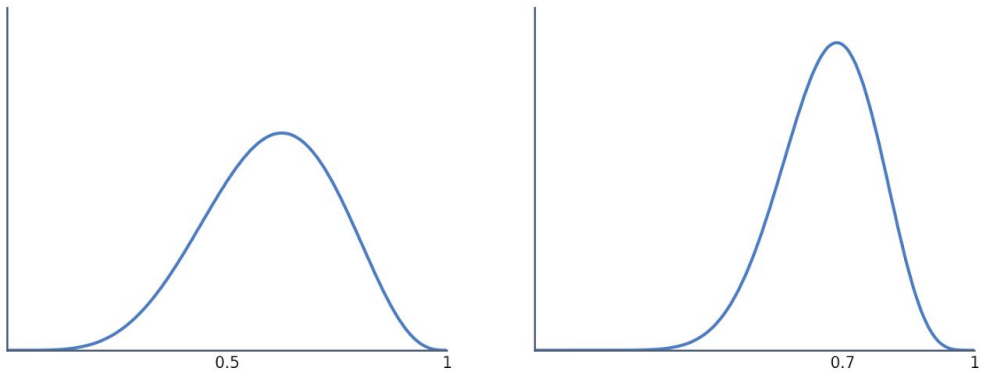

Fig. 2: The probability distributions following conditionalization

The mean of the distribution in the left panel shifts from 0.5 to 0.6, while that of the distribution in the right panel shifts from 0.7 to 0.66. This is because in the left panel, we started with an attitude of weak indifference between all hypotheses, and we updated that attitude on a sequence of experiments in which a slight majority of heads was observed. Thus, the experiments nudged us up toward expecting that the coin is biased toward Heads. Meanwhile, in the right panel, we started with a strong hunch that the coin is significantly biased toward Heads, which we updated on a sequence of experiments that is only slightly dominated by Heads. Thus the experiments nudged us down toward suspecting a less extreme Heads bias. Thus the two agents' credences change by different amounts, and in different directions, in a way that reflects updating by Bayesian conditionalization on the evidence observed.

Gender Bias Study is analogous to the coin toss example. There is an unknown parameter, analogous to the unknown bias of the coin: the actual proportion of women employed at universities who are faculty as opposed to administrative assistants. Before reading about the study, you have no information about the value of this unknown parameter. But you know that it must have some value between 0 (0% faculty) and 1 (100% faculty), just as Boris knows the range of possibilities about the coin's bias. So you can distribute your credence over hypotheses about the value of this unknown parameter, just as Boris does. This prior probability distribution will correspond to a pair of numbers representing pseudo "observations" — to put things less oddly (if it is any less odd) we might think of them as hypothetical prior encounters — of women employed at universities who are faculty and of women employed at universities who are administrative assistants. When you read about the study, you gain data to use in your estimate of the value of the unknown parameter. Updating by conditionalization will then move you to the posterior probability distribution corresponding to the pair of numbers that result from adding the number of women faculty reported in the study to your pseudo "observations" of women faculty and adding the number of women administrative assistants reported in the study to your pseudo "observations" of women administrative assistants. As with the coin toss, at any point in time your predictive estimate of the probability that the next woman employed at a university whom you meet will be a faculty member is the mean of your probability distribution at that time.

This is already enough to see why Gendler's sad conclusion is false. But let us explain more carefully why this is the case.

The reasoning behind the sad conclusion depends on this principle:

Frequency-Credence Connection: If (a) I know that a is an F, and (b) I know that x% of previously sampled Fs are G, and (c) I have no further evidence bearing on whether a is G, then my credence that a is G should be x.

This is a popular principle. Contemporary formal epistemologists continue to defend it and apply it to cases like **Gender Bias Study** (see e.g. White 2010, Buchak 2013). But the principle has roots in some highly influential early work in probability theory (see e.g. Reichenbach 1938, 1949; Kyburg 1974). The frequency-credence connection states that we should conform our credences to observed frequencies when making predictive inferences; for example, that we

should have credence 0.3 that Mary will be a faculty member based on the reported results of the study.

The frequency-credence connection cannot be a genuine principle of epistemic rationality. This is because, given the way predictive inference works, agents can only adopt the posterior credence that this principle recommends if they had a prior credal distribution that violates the probability axioms. Posterior credences usually depend on two things: an observed frequency, and the agent's prior credences. But the frequency-credence connection requires that an agent's posterior credence exactly equal the observed frequency. This principle thus requires that her prior credences have no influence on her posterior credences; in other words, it requires her to behave as if she had no prior. Of course, an agent cannot actually have no prior – or, if she did, then she could not engage in Bayesian conditionalization at all, as she would have nothing to update. Now, it turns out that there is one prior credal distribution that exerts no weight on the posterior and thus allows agents to behave as if they had no prior. However, for mathematical reasons that we will not go into here, there is only one of these, and it looks quite strange: its graph is a U-shaped parabola over the interval [0,1], which increases arbitrarily as it approaches the extremal credal values of 0 and 1.[1] The area under this parabola is infinite, since the graph approaches 0 and 1 only at the limit. But this violates the probability axioms. For the axioms require that an agent's credences sum to 1, not infinity. Such a prior is therefore irrational.

Some credences that do not sum to 1 can be saved from irrationality using a standard trick in probability theory: we can transform incoherent credences into coherent ones using a *normalizing constant* by which we multiply all the credences. For example, if someone's credences in the elements of a partition sum to 3 rather than 1, then we can "normalize" them by multiplying them all by 1/3. But the "U"-shaped prior that a Bayesian agent would have to adopt in order to respect the frequency-credence connection cannot be rescued in this way, since there is no real number such that when we divide infinity by that number we get 1. (i.e. there is no reciprocal to infinity.) Thus, the prior credal distribution that agents would need to adopt in order to respect the frequency-credence connection requires them to be irremediably probabilistically incoherent. This sort of reasoning is therefore prohibited even by the minimal principles of epistemic rationality that accuracy-based epistemology endorses.

So, how should we make predictive inferences?

On our accuracy-based conception of the requirements of epistemic rationality, this question does not have a single correct answer. That is because no single prior credal distribution is uniquely required by epistemic rationality. Rather, many such distributions are epistemically permitted, and different agents may adopt different ones according to their attitudes toward *epistemic risk*.

Someone's attitude toward epistemic risk reflects the way she thinks about the two types of epistemic error: high credence in a falsehood and low credence in a truth. One of the

[1] We can also think informally about how weird this prior is. Someone with this prior is virtually certain that the value of the unknown parameter is not 0.5, and is infinitely confident both that it is somewhere close to 1 and that it is somewhere close to 0. That is very weird.

fundamental tenets of accuracy-based epistemology is that high credences in falsehoods and low credences in truths are bad. But this says nothing about the relative badness of the two types of error. In formal circles, one often hears that the platitude that we should pursue the "Jamesian goals" of believing truths and disbelieving falsehoods does not tell us how to prioritize these goals. Thus, someone may think that approaching error in the *false positive direction* — by increasing one's credence in a falsehood — is always equally as bad as approaching error in the *false negative direction* — by decreasing one's credence in a truth — if the increase and decrease are in equal amount. But someone may equally well think that one of these types of error is worse than the other. And someone might think that which type of error is worse (if any) depends on the proposition in question. For example, weather forecasters may reasonably be more worried about low credence in truths than high credence in falsehoods if the proposition in question is that a tornado is nearby, since a false alarm would be inconvenient but failing to predict a tornado would be disastrous. In this case, the forecaster understands that false negatives are worse than false positives. Moreover, as this example shows, we assess the disvalue of the two types of epistemic error based on non-alethic concerns. The forecaster does not have an intrinsic aversion to false negatives. Rather, her asymmetric attitudes to the costs of error are based on her understanding of the practical costs involved in the case. But this is fully compatible with the requirements of epistemic rationality on an accuracy-based approach; it does not lead her to violate any of the aforementioned requirements.

Following Babic (2019), we can model agents as having an *epistemic risk function*, whose inputs are credences in propositions and whose outputs are numbers representing the agent's assessment of the "riskiness" of this credence in this proposition. Whenever anyone assigns a credence to a proposition, she "risks" being penalized with a certain inaccuracy score if her credence turns out to be far from the truth — that is, a low credence in a falsehood or a high credence in a truth. Thus different credences will seem more or less "risky" to different agents, depending on their view of the relative disvalue of the two types of epistemic error. For example, the weather forecaster regards errors in the false negative direction as much more serious than errors in the false positive direction for the proposition that there is a tornado nearby, so she will see lower credences as "riskier" than higher ones. She thus evaluates epistemic risk *asymmetrically*. By contrast, someone who sees the two types of error as equally bad will evaluate epistemic risk *symmetrically*. This means that she will assess low and high credences as equally risky, provided that they deviate from 0.5 by an equal amount. For example, she will see credence 0.2 and 0.8 as equally risky.

In Babic (2019), one of us has shown that epistemic risk functions meeting a handful of intuitive criteria are uniquely associated with monotonic, continuous, and strictly proper scoring rules (Babic 2019). This result suggests that, since the requirements of epistemic rationality are not specific enough to pin down a single correct scoring rule, the agent may choose among them based on her attitude toward epistemic risk.

An agent's attitude toward epistemic risk can also rationalize her choice of a credal distribution over hypotheses about the value of an unknown parameter — like the bias of a coin, or the proportion of women employed at universities who are faculty — before she gains any evidence about it. This is because, before gaining evidence about some proposition, agents can prefer

some credences to others on the basis that they *minimize epistemic risk*. The least risky credence is the one that guarantees the agent a certain inaccuracy score no matter whether the proposition in question turns out to be true or false.

To see how this applies in cases like **Gender Bias Study**, consider someone whose way of thinking about predictive inference seems morally good: someone who holds that mistakes in the false negative direction are worse than mistakes in the false positive direction when the proposition in question is that Mary is a faculty member (call this proposition 'm'). For clarity, suppose that this person's epistemic risk function is as follows, with credences in the proposition that Mary is faculty along the *x*-axis and her assessment of those credences' riskiness along the *y*-axis:

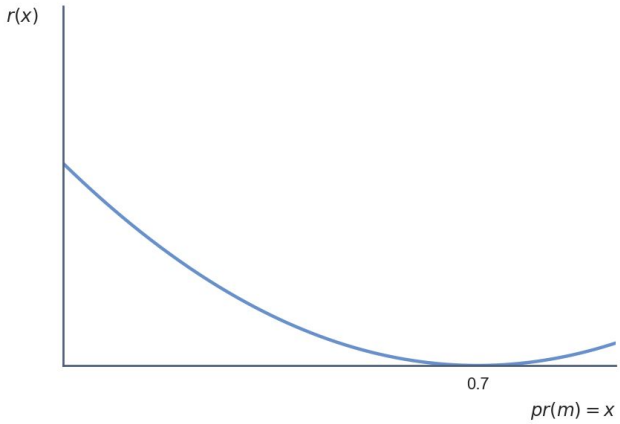

Fig. 3: An asymmetric epistemic risk function

This attitude toward epistemic risk can help the agent to choose a prior credal distribution over hypotheses about the proportion of women employed at universities who are faculty before she reads about the study. For someone who assessed the costs of error symmetrically, her prior credal distribution would have to be the uniform distribution depicted in the left panel of Figure 1. This is because, for such a person, the unique credence that minimizes epistemic risk is 0.5, and the uniform distribution is the only one with this mean. But our agent has more options. Given the epistemic risk function depicted in Figure 3, the minimally risky credence is 0.7 — this is the credence such that the agent is guaranteed the same inaccuracy score whether the proposition in question turns out to be true or false. Thus our agent should adopt a prior credal distribution whose mean is 0.7, so as to minimize epistemic risk. But very many credal distributions have this mean. For example, one is the distribution corresponding to 100 pseudo "observations" of women administrative assistants and 300 pseudo "observations" of women faculty. So our agent might adopt this prior distribution over possible values of the unknown parameter, justified by her attitude toward epistemic risk. Such a prior distribution would look like this:

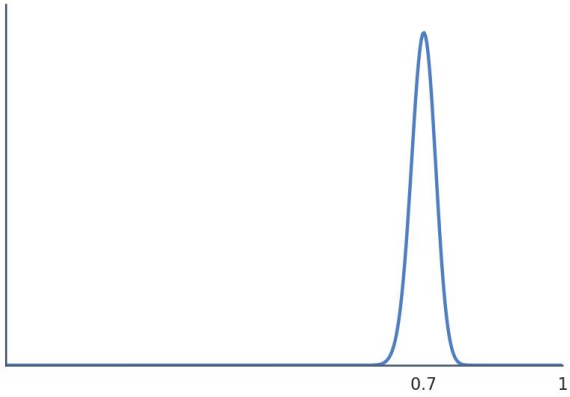

Fig. 4: Probability distribution licensed by the attitude toward epistemic risk depicted in Fig. 3

Now consider what happens when our agent reads about the study. The newspaper reports that the study "observed" 350 women employed as administrative assistants and 150 employed as faculty. If our agent updates by Bayesian conditionalization, then she will move to the posterior distribution corresponding to 100+350 encounters with women administrative assistants and 300+150 encounters with women faculty. Her prior (light blue) and posterior (dark blue) distributions are represented in the right panel of Figure 5. For comparison, the left panel shows a pair of possible prior and posterior distributions for someone who sees the two types of error as almost equally disvaluable, in response to the exact same evidence; this person begins with a weak prior centered around 0.5, and when she updates by conditionalization she will move to a credal distribution that is approximately equal to 350 encounters with women who are administrative assistants and 150 encounters with women who are faculty.

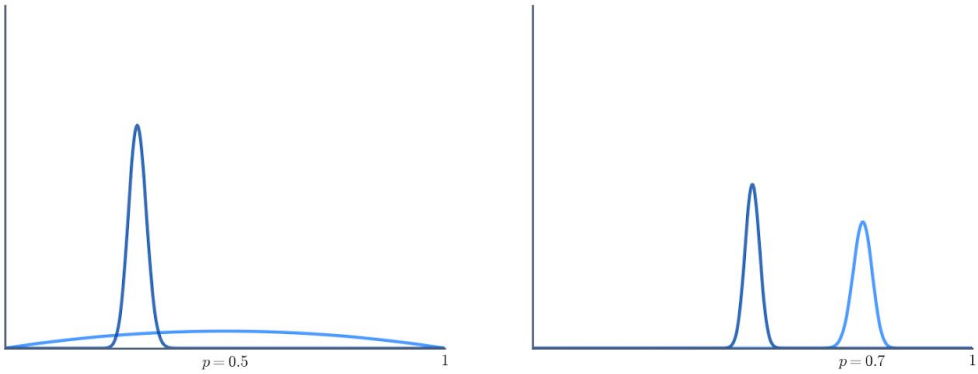

Fig. 5: Two probability distributions before (light blue) and after (dark blue) conditionalization

Here the person who thinks that the costs of error are symmetric will end up, after conditionalizing on the results of the study, with a credal distribution such that she thinks Mary

is very unlikely to be faculty. The mean of her distribution is now 0.31, so that will be her credence that Mary is a faculty member — not far from the 0.3 recommended by the Frequency-Credence Connection. This person thus displays precisely the sort of epistemic behavior that seems to wrong Mary. But that is not so for the person who thinks that the costs of error are asymmetric. The mean of her credal distribution after conditionalizing on the results of the study is 0.5, so she will judge Mary exactly equally as likely to be faculty as she is to be an administrative assistant.

To emphasize: our agent depicted in the right panel of Fig. 5 is doing just as well, epistemically speaking, as the person who assesses the costs of error symmetrically and displays the updating behavior shown in the left panel. Our agent's credences obey the probability axioms throughout. She updates by conditionalization. And she assesses expected accuracy using a monotonic, continuous, and strictly proper scoring rule. Thus, she meets all of the requirements of epistemic rationality. But our agent avoids leaping to the conclusion that Mary is probably an administrative assistant solely on the basis of the information in the study. Thus, she avoids doing what is morally problematic, without doing anything that is epistemically prohibited. She shows that it is indeed possible to be both rational and equitable. She shows, then, that the Sad Conclusion is false.

Many formal epistemologists who endorse the frequency-credence connection would disagree with us on this last point. This is because they think that, before someone gains any evidence about which of a set of hypotheses is true, epistemic rationality requires that she be "indifferent" between these hypotheses — where this "indifference" is construed as a uniform distribution like that in the left panel of figure 1. (On this see e.g. White 2010). One often hears that a uniform distribution represents indifference because it is "neutral" as to the truth or falsehood of the proposition in question. But that is a mistake. The uniform distribution is not neutral at all. On the contrary, a uniform distribution corresponds to a precise number of pseudo "observations", just as any other distribution does. And a uniform distribution reflects a maximally specific attitude toward the relative costs of the two types of epistemic error, just as any other distribution does: this distribution corresponds to an epistemic risk function according to which the two types of epistemic error are exactly equally bad.

To be clear: we agree that the uniform distribution is epistemically permitted. We deny that it is epistemically privileged. Someone may hold this attitude toward epistemic error, but she had better have an argument for it, since there is excellent reason to think that the two types of error are *not* exactly equally bad in this case — viz., all the moral and political reasons to think that one type of error is significantly worse than the other. Moreover, there are epistemically permitted attitudes toward epistemic risk that take proper account of these moral costs. If someone eschews these all-things-considered normatively attractive epistemic options, and instead insists on assessing the costs of error symmetrically, then we hold that the problem with this person is not that they are epistemically irrational but simply that they are a jerk.

It may be tempting at this point to object on the grounds that bringing moral and political costs into the assessment of epistemic rationality is inappropriate. But our point is that there is no way to avoid this. One cannot give a purely alethic story for why the two types of error are to be

treated equally; this is what epistemologists acknowledge when they talk of different ways to balance the "Jamesian goals". In other words, to claim that the two types of error are equally costly is not to avoid taking a moral or political position — it is itself a moral or political position. While it may look like the "default" option, it is not. In the absence of any information, a uniform distribution is no more plausible than any other maximally specific distribution.

At this point we should clarify a possible misconception. We are not saying that it is epistemically permissible for someone's attitude toward epistemic risk to be such that she will *never* reach the point where she thinks that the next member of a certain social group whom she meets is more likely than not to possess a certain socially disvalued trait. We have not argued that every possible attitude toward epistemic risk is epistemically permissible. Nor, equivalently, do we hold that every possible prior probability distribution is epistemically permissible. Rather, we hold that a wide range of attitudes toward epistemic risk are epistemically permissible, and that the same goes for a corresponding range of probability distributions. This range includes several attitudes and priors that do not lead the agent to draw the pernicious predictive inference in **Gender Bias Study**. What exactly the bounds of this range are remains an open question that we do not intend to settle here. And, for each permissible attitude toward epistemic risk, there is a quantity of evidence that the agent could in theory obtain that would shift her probability distribution over possible values of the unknown parameter to the point where its mean is above 0.5. So, our picture of epistemic rationality does not instruct agents to simply ignore their evidence for moral and political reasons. Rather, for each agent, there may come a point where she will take the next group member whom she meets to be more likely than not to possess the relevant socially disvalued trait. But our morally virtuous agents reach this point more slowly than the "neutral" ones, and it takes more evidence to get them there. Their attitudes toward the differential costs of false positive and false negative errors ensure that they need a lot more evidence than others do before they begin to expect members of a certain social group to bear socially disvalued traits solely on the basis of information about the prevalence of those traits within the group.

One might take this last point to show that the requirements of epistemic rationality and those of morality do inevitably conflict after all. But we think that this would be a mistake. That is because it is unclear precisely what morality requires in cases like **Gender Bias Study**. We stipulatively introduced the label "pernicious predictive inference" above, but it is far from clear that all inferences of this type are pernicious, and in some cases they may be morally required. This is because, while there are good moral reasons to be reluctant to assume that Mary is an administrative assistant based solely on the information in the study, there are further moral reasons *not* to exhibit such reluctance, or at least to limit its scope. For one thing, reluctance to assume that people are administrative assistants reinforces the idea that there is something bad about being an administrative assistant. There are moral reasons to avoid assuming that women and minorities hold low-status jobs, but there are also moral reasons to question the status of these jobs, affording so-called "women's work" its proper value. This generalizes; many socially disvalued traits that are statistically prevalent in marginalized social groups are disvalued (at least in part) *because* of their association with these groups. In such cases, we have moral reason not to act as if possessing the trait is a bad thing, and thus not to exhibit reluctance to draw predictive inferences about it. Moreover, there is something sanctimonious and disingenuous

about refusing ever to expect any members of any marginalized groups to bear any socially disvalued traits, regardless of what we learn about the prevalence of the traits within the groups. Just as "color-blind" policies can make us overlook the contemporary effects of historical racial injustices, similarly this extreme predictive reluctance can make us fail to respond adequately to the fact that someone belongs to a social group in which a disvalued trait is highly prevalent, ignoring the ways in which this fact impacts the person's life. So it seems to us highly likely that, in each real-life case, morality calls for a less-than-maximal degree of predictive reluctance. We assume that the attitude toward epistemic risk that yields this morally optimal degree of predictive reluctance falls within the range of epistemically permissible attitudes toward epistemic risk.

3. Other Types of Statistical Inference

This paper has so far focused on predictive inference: an assessment of the probability that an individual possesses a certain trait, based on information about the trait's prevalence among previously observed members of a group to which that individual belongs. But this is not the only kind of inference someone can perform on the basis of information about the prevalence of a socially disvalued trait within a certain social group. In cases like **Gender Bias Study**, the agent is using information about observed frequencies among other group members to make a prediction about a new group member whom she encounters. In another kind of statistical inference, called *direct inference*, the agent makes a prediction about someone in the very sample from which the observed frequency was taken. For example, **Gender Bias Study** would be a case of direct inference if you knew that Mary was one of the people interviewed for the study.

In some cases of direct inference, the agent's credence that the encountered group member possesses the trait is, after all, epistemically required to be equal to the observed frequency of the trait within the group. This is so under conditions of *exchangeability*, in which the agent has no information about group members that makes any of them more or less likely to possess the trait than any others. But when it comes to real-life cases of direct inference, such conditions are extremely rare. In all of the real-life cases that worry philosophers and legal scholars, conditions of exchangeability do not hold. That is because the agents in these cases all encounter people in such a way as to learn a lot of information about them that either raises or lowers the probability that they possess the trait. For example, in **Gender Bias Study** you not only learn that Mary is a woman, but that she lives in your area and has a certain appearance. This means that she is not just as likely to be a faculty member as any other woman from the study, based on your information. Since genuine conditions of exchangeability are vanishingly rare in real-life cases, we are not worried about direct inference. The circumstances in which someone could be epistemically required to conform her credences to observed population frequencies are possible, but extremely unlikely to arise.

Another thing one can do based on information about the prevalence of a socially disvalued trait within a certain social group is *accept the statistic*. For instance, in **Gender Bias Study**, you can accept that 70% of women employed at universities like your local university are administrative assistants and only 30% are faculty. Or you can accept that these were the proportions found in a sample of such women. Or, if you are in a more skeptical mood, you can accept that a study

found that these were the proportions in a sample of such women, or that a newspaper reported that this is what the study found.

There is no epistemic requirement to assume that the overall frequency of a trait in a population is equal to the frequency that one hears that a study found in a sample of that population. (This is why statistical information about the prevalence of a trait within a sample can inform an agent's probability distribution over possible values of the actual population frequency — the "unknown parameter'' discussed above — but does not dictate precisely what her estimate of this value should be.) So there is no epistemic requirement to accept statistics in this sense. But some related epistemic requirements do hold. The agent in **Gender Bias Study** cannot simply ignore the information that she hears, revising none of her doxastic states. On the contrary, she has gained new evidence, and she must decide what her evidence is and respond accordingly. If there is reason to doubt the study's validity (e.g. because it used a small sample, or excluded observations that should not have been excluded) then she may take her evidence to be that *a study found* evidence of a certain population frequency rather than that *there is* a certain population frequency. And if there is reason to doubt the newspaper's reliability (e.g. because it has a track record of inaccurately reporting scientific studies) then she may take her evidence to be that *a newspaper reported* that a study found evidence of a certain population frequency. Similarly, if she has reason to be even more skeptical, then she might take her evidence to be that *she had a series of sensory experiences as of* a newspaper reporting that a study found evidence of a certain population frequency. But she must start somewhere. This is how responding to evidence works; we update our credences by conditionalizing on what we have learned, and conditionalization begins with the agent identifying something that she has learned.

Again, one might think that this point supports the Sad Conclusion after all. But that would be a mistake. This point supports the Sad Conclusion only if there are statistics that we are both epistemically required to accept and morally prohibited from accepting. And, again, these appear to be vanishingly rare. Indeed, when statistics about population samples from well-designed studies are reported by credible sources, accepting them may be morally required. This is because we are morally required to resist the structural injustices that underlie and explain these statistical facts, and we cannot do so effectively without informing ourselves of the facts. For example, we cannot address the structural problems underlying incarceration rates among contemporary African-Americans without accepting information about what those rates are. Similarly, when there is reason to distrust the source of a piece of statistical information, someone may still be morally required to accept that the source reported the information. For example, we cannot address the structural problems underlying incarceration rates among contemporary African-Americans if we ignore information about how these rates are reported, particularly if public reports are systematically misleading. Thus the Sad Conclusion is, again, false: there is no tension between the requirements of epistemic rationality and those of morality. Rather, the sort of epistemic behavior that might be rationally required — accepting data about the prevalence of pernicious traits within samples of social groups, or information about how such data is reported — is at least morally permitted, and may sometimes be morally required.

4. Conclusion

We have argued that Gendler's "sad conclusion" is false: there is no inevitable conflict between the requirements of epistemic rationality and those of morality, on a standard Bayesian construal of the requirements of epistemic rationality. We agree with Gendler that the "neutral" attitude (corresponding to a uniform prior) is criticizable on moral and political grounds. We do not think that it is criticizable on epistemic grounds. But we have argued that there are a family of attitudes toward epistemic risk that are not criticizable on either grounds. These are attitudes that render the agent slow to draw pernicious predictive inferences, while remaining consistent with the requirements of epistemic rationality. So there are plenty of all-things-considered normatively attractive options in cases like **Gender Bias Study**; that is, plenty of options that allow us to combine being rational with being equitable. This means that prejudiced people cannot claim that cold, calculating epistemic rationality is on their side; properly understood, it is easy to see that it isn't.

We stated at the outset that our approach is intended to complement, rather than challenge, existing ways of avoiding the Sad Conclusion. But there is something positive to be said for our approach. As we noted earlier, it is striking that existing approaches all attempt to secure the strong result that, in at least some cases like **Gender Bias Study**, the kind of predictive inference that seems morally unsavory is also epistemically prohibited. To secure this strong result, some existing approaches apply only to certain cases of pernicious predictive inference — those in which the agent makes a particular epistemic error. And others apply to all cases, but do so by incorporating a substantial amount of contentious theoretical machinery that complicates our understanding of the requirements of epistemic rationality. Meanwhile, we are able to vindicate the *option* to be slow to draw pernicious predictive inferences in cases like **Gender Bias Study**, by showing that, as well as being morally desirable, such predictive behavior need not be epistemically irrational. And our approach secures this weaker result by locating the point at which moral, political, and other practical considerations already factor in to a simple accuracy-based understanding of the requirements of epistemic rationality: these considerations are the basis for agents' assessments of the relative costs of the two types of epistemic error, thus determining her attitude toward epistemic risk. This means that our approach applies in all cases and requires no contentious additional theoretical machinery. We take this to be a significant advantage of our approach: it is (arguably) the most minimal, and thus (hopefully!) the least controversial, way of avoiding the Sad Conclusion.

Bibliography

Babic, Boris. 2019. "A Theory of Epistemic Risk." *Philosophy of Science* 86 (3): 522-550.

Basu, Rima. and Schroeder, Mark. 2019. "Doxastic Wronging," in *Pragmatic Encroachment in Philosophy*, eds. Brian Kim and Matthew McGrath. Routledge: 181-205.

Basu, Rima. 2019. "What We Epistemically Owe to Each Other." *Philosophical Studies.* 176 (4): 915-931.

Buchak, Lara. 2014. "Belief, Credence, and Norms." *Philosophical Studies*, vol. 169 (2): 285–311.

Cohen, Jonathan. 1981. "Subjective Probability and the Paradox of the Gatecrasher." *Arizona State Law Journal*, vol. 1981 (2): 627–634.

Enoch, David, Levi Spectre & Talia Fisher. 2012. "Statistical Evidence, Sensitivity, and the Legal Value of Knowledge." *Philosophy & Public Affairs*, vol. 40 (3).

Gardiner, Georgi. *ms*. "Evidentialism and Moral Encroachment".

Gendler, Tamar Szabo. 2011. "On the Epistemic Costs of Implicit Bias." *Philosophical Studies* 156 (1), 33-63.

Greaves, Hilary and David Wallace. 2006. "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility." *Mind* 115(459), 607-632.

James, Wiliam. 1896. "The Will to Believe." *The New World* 5, 327-347.

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

Kaplan, John. 1968. "Decision Theory and the Factfinding Process." *Stanford Law Review*, vol. 20 (6): 1065–1092.

Kyburg, Harry. 1974. *The Logical Foundations of Statistical Inference*. Dordrecht: Reidel.

Munton, Jessie. *ms*. "The Epistemic Flaw with Accurate Statistical Generalizations"

Moss, Sarah. 2016. *Probabilistic Knowledge*. Oxford, UK: Oxford University Press.

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

Reichenbach, Hans. 1949. *A Theory of Probability.* Berkeley: University of California Press.

Thomson, Judith Jarvis. 1986. "Liability and Individualized Evidence." *Law & Contemporary Problems*, vol. 49 (3): 199–219.

Tribe, Laurence H. 1971. "Trial by Mathematics: Precision and Ritual in the Legal Process." *Harvard Law Review*, vol. 84 (6): 1329–1393.

White, Roger. 2010. "Evidential Symmetry and Mushy Credence." In T.S. Gendler and J. Hawthorne (eds), *Oxford Studies in Epistemology*, Vol 3, 161-186. Oxford University Press.