Philosophers'
Imprint

volume 23, no. 12
july, 2023

Moral Encroachment under Moral Uncertainty

Boris Babic, Zoë Johnson King

University of Toronto, Harvard University

Abstract

This paper discusses a novel problem at the intersection of ethics and epistemology: there can be cases in which moral considerations seem to "encroach" on belief from multiple directions at once, and possibly to varying degrees, thereby leaving their overall effect on belief unclear.[1] We introduce these cases – cases of *moral encroachment under moral uncertainty* – and show that they pose a problem for all predominant accounts of moral encroachment. We then address the problem by developing a modular Bayesian framework that, we argue, is sufficiently flexible and scaleable to accommodate the multifaceted uncertainty that we describe while still generating clear recommendations. Our framework has several practical upshots, so we close by articulating them: we examine the relationship between moral character and doxastic behavior and make suggestions for how to encourage people to revise their doxastic states in morally laudable ways, without deviating from core Bayesian norms.

1. Introduction

This paper develops a modular Bayesian framework for thinking about moral encroachment under moral uncertainty. To our knowledge, it is the first paper to explicitly consider and develop a way for thinking about moral encroachment and moral uncertainty in tandem.

There is a large and growing literature on moral uncertainty, concerning how we should decide what to do when we are unsure of the moral facts. For example: Joy must decide between eating a burger or a salad. In addition to her uncertainty about the non-moral facts relevant to her decision – What will each dish taste like? Which one will she enjoy more? Are they cherry tomatoes or regular tomatoes in the salad? – she may also face uncertainty about the relevant moral

[1] Both authors contributed equally to the writing of this paper at all stages of development.

facts – Is consuming beef morally permissible? What is more important, promoting animal welfare or lowering one's "food miles"? Does her preference for one dish over the other based on taste have any moral weight, and if so, how much weight does it have? The literature on moral uncertainty concerns the principles that should govern decisions like Joy's. Likewise, there is a robust and active debate about moral encroachment, concerning whether moral facts impinge on epistemic rationality and, if they do, how exactly they do so. For example: Adela learns that in four out of five cases of maple syrup theft, a Canadian is the thief. Should she blame Lawrence, the only Canadian suspect, on the basis of this statistical evidence? Or would there be something morally amiss with blaming Lawrence on this basis and, if so, does that affect the epistemic credentials of a belief that he is guilty? Another example: Xiangyuan thinks that he remembers locking the door to his grandmother's house, but then he sees a police report about walk-in burglaries in her area. Does the fact that it would be terrible for his grandmother to suffer a burglary mean that his vague memory of having locked the door no longer suffices for justification? The literature on moral encroachment concerns the doxastic attitudes of people like Adela and Xiangyuan.[2]

We are interested in situations where these two phenomena occur together. For example, suppose that one of your students tells you that one of her classmates sexually assaulted her. You must decide whether, or to what extent, to believe your student. Now, there is a norm with a distinguished history in anglophone jurisprudence suggesting that you should be highly cautious about blaming or accusing someone of criminal behavior because, the thought goes, it's much worse to punish an innocent person than to excuse a guilty one. Channelling the legal literature, we'll call this norm the *Benefit of the Doubt Norm*. Were this the only relevant norm, it would be clear how it bears on your beliefs: it urges caution against confidence in the proposition that the classmate committed the sexual assault. However, there is a competing norm that recommends giving preferential epistemic treatment to the victim instead, stemming from the reality that women's sexual harassment and assault complaints often go unanswered and that rectifying this injustice requires epistemic boldness rather than caution. We'll call this norm the *Victim Deference Norm*. If the Victim Deference Norm stood alone in this situation, its upshot would be likewise clear: it encourages confidence that your student is telling the truth.

We use these norms as illustrative toy examples. Their precise content is not important to our argument; rather, the point we want to highlight is that in cases like this there are competing moral considerations at play. What makes this case particularly interesting to us – and what makes it a case of moral encroachment under moral uncertainty – is that, while you may have some thoughts about the relative importance of these competing moral norms, most of us are not sure of *precisely* how they fare with respect to each other.[3] As a result, it is not immediately obvious how they ought to jointly impinge on belief. Should we give more weight to the Benefit of the Doubt Norm? Or the

[2] Here, and throughout the paper, we use "doxastic attitude" to refer neutrally to both beliefs and credences. We will develop a credal model. But many authors in the existing literature on moral encroachment approach the problem in terms of full beliefs instead. Indeed, some authors have assumed that moral considerations encroach on belief but *not* on credence (Bolinger (2020a), Fritz (2020), Gardiner (2018)), whereas others have defended moral encroachment on credences (Fritz and Jackson (MS), Johnson King and Babic (2020)). We remain neutral for now, flagging differences between the two approaches insofar as they are relevant. But we will ultimately adopt an approach that assumes moral considerations do encroach on credences and will develop a model of the way in which they do so.

[3] Indeed, most of us are not even sure of *precisely* what each of these two norms, considered individually, requires of us. It should be apparent that the norms cannot plausibly be understood as exceptionless rules; it isn't plausible that you should *always* be highly cautious about blaming or accusing someone, nor that you should *always* give preferential epistemic treatment to a(n alleged) victim. This uncertainty about how to understand each norm could quite reasonably compound your uncertainty about the relative moral importance of the two norms in the case at hand. We will say a little more about this later in the paper.

Victim Deference Norm? Or do they cancel each other out, so to speak? Or should we apply a more fine grained trade-off, say 2:1?

In this paper, we develop a normative framework for thinking about such problems. In doing so, we draw on and generalize the Bayesian encroachment view developed in Johnson King and Babic (2020) and Babic et al. (2021). In addition to motivating and constructing a flexible model for thinking about moral encroachment under moral uncertainty, we also draw some insights from our model with respect to different types of moral character and doxastic behavior: we will show that people who are maximally morally uncertain often act like people who are indifferent to the potential moral costs of their beliefs and behavior. This suggests that, while open-mindedness usually seems like an epistemic virtue, in cases of moral uncertainty it can turn to vice – a vice we call *moral spinelessness*. As a result, in cases like the above it seems one ought to be somewhat bold. However, we will also show that there are in fact two ways to increase the influence of one's moral convictions on one's belief: one can achieve this either by being sufficiently bold with respect to moral uncertainty, or by being sufficiently skeptical about one's evidence. We take it to be a virtue of our model that it illuminates this interplay between moral convictions and doxastic behavior.

The paper proceeds as follows. In Section 2, we describe the basic structure of problems involving moral encroachment under moral uncertainty. In Section 3, we summarize the existing literature and demonstrate that approaches to moral encroachment that do not accommodate moral uncertainty (which includes all existing approaches) invariably fall short. In Section 4, we describe our model in full and explain how it provides a useful lens for reasoning about cases of moral encroachment under moral uncertainty. Finally, in Section 5, we discuss the implications of our model, focusing in particular on different attitudinal responses to moral uncertainty. The value of this paper lies in its describing a new problem for the literature on moral encroachment, developing a formal model that can handle that problem, and drawing surprising normative insights from the model with respect to how one ought to regulate one's doxastic states and encourage others to regulate theirs.

2. Setup

To begin, consider the following example:

> **Sports Store.** You are in a sports apparel store where the floor employees do not wear uniforms. Instead, they wear the apparel sold at the store. This can make it difficult to tell employees apart from other customers, many of whom wear the same or similar sportswear. You are looking for an employee. You see a young person of color, dressed in sportswear, hanging clothing items back on the racks. What should be your doxastic attitude toward the proposition that this person is an employee?

This example illustrates several features in which philosophers are interested when discussing moral encroachment.[4] First, you have plenty of evidence that the person before you is an employee, but some pieces of your total evidence seem more problematic than others. You have (let's assume) some background statistical evidence suggesting that this person belongs to a social demographic – young people of color – that comprises a relatively high proportion of sports store employees, due to underlying sociopolitical, cultural, and economic factors. But you should remain mindful that individuals can be exceptions to accurate statistical generalizations (Moss, 2018a). And the person in front of you might seem to have a moral complaint against your assuming that they are an employee based on demographic profiling, which they do not

[4] A structurally similar example – the now well-known Cosmos Club example – appears in Gendler (2011) and has since been widely discussed.

similarly have if your assumption is instead based on the observed behavioral evidence that they are wearing some of the sportswear sold at the store and hanging clothes on a rack (Bolinger, 2020a). Moreover, there is a salient alternative possibility that none of your evidence rules out: that this person is a customer who just tried on some clothes, decided not to buy them, and is helpfully re-racking them. Since you cannot rule this alternative out, a belief or high credence that the person before you is an employee may not amount to knowledge (Moss, 2018a). Indeed, even if this alternative possibility has not occurred to you, the moral costs involved in erroneously assuming that the person is an employee might suffice to render the error possibility relevant and thus – since you cannot rule out a relevant alternative – to prevent your doxastic state from constituting knowledge (Moss, 2018b).

Indeed, the moral costs of this type of epistemic error may mean that you do not have enough evidence to conclude that the person in front of you is an employee. In terms of full beliefs, this could be because the high moral "stakes" raise the bar at which evidence suffices for justified belief to a point that your current evidence fails to meet (Fritz (2017); cf. Worsnip (2021)). In terms of credences, the high moral costs of a false positive error may mean that you should increase your credence in the proposition that the person in front of you is an employee more slowly than you would for a proposition that is not similarly morally "risky", with the result that your posterior probability remains below 0.5 even after updating on your evidence (Johnson King and Babic, 2020).

One might also think that a belief or high credence that this person is an employee based on your evidence is morally costly *regardless* of its truth or accuracy. Perhaps, whether or not it is true that the person in front of you is an employee, adopting a belief or high credence in this proposition based (at least in part) on your statistical evidence would wrong them by taking an objectionably "clinical" or predictive attitude toward them, thereby diminishing their agency.[5] And perhaps this direct moral cost is itself sufficient to undermine your doxastic state's epistemic status, even setting aside the further moral risks associated with the possibility of error (Basu, 2019a,b; Basu and Schroeder, 2019; Fritz and Jackson, MS).

We will say that a moral encroachment *theorist* is someone who thinks that moral considerations can affect the epistemic status of some doxastic states without bearing directly on the truth or accuracy of those states. Someone who accepts any of the views just sketched counts as a moral encroachment theorist for our purposes. By contrast, we will say that a moral encroachment *practitioner* is someone who is convinced that a view somewhere in this vicinity is correct and would like to put this view into practice when forming, maintaining, and revising their own doxastic states.

It is fairly clear what a moral encroachment practitioner should do in a case like **Sports Store**. If they are confident that there are either moral risks or direct moral costs associated with assuming that the person in front of them is an employee, but no parallel moral risks or costs associated with assuming that this person is a fellow customer or with remaining agnostic, then they should adopt one of the latter attitudes. This could be because they want to avoid doxastically wronging the person in front of them. Or it could just be because they want to avoid having a doxastic state that fails to constitute knowledge, either due to their inability to eliminate error possibilities or due to the inadequacy of their evidence in light of the moral stakes. Regardless of which argument moves them, it is fairly easy for the moral encroachment practitioner to practice moral encroachment in this case: they just hold off from assuming that the person in front of them is an employee. For credences, things are a little less obvious, but there are still some

[5] One might think this, for instance, if one holds that treating someone as an autonomous agent prohibits taking the Strawsonian "objective stance" toward them – an idea to which we will return in the next section.

relatively straightforward options: you can update in a manner that is sensitive to the moral risks and thus revise your credences more slowly for more morally risky propositions than for less risky ones (Johnson King and Babic, 2020), or you can bracket some of your evidence entirely when a morally costly proposition is at stake, refusing to update on it at all (Fritz and Jackson, MS).[6]

All of the above holds in any case in which the moral risks or costs are *all on one side*, so to speak; that is, in which moral risks or costs attach to believing but not to disbelieving or withholding, or to adopting one extremal credence (i.e. an extremely high or low credence) but not the other.

But the moral risks are not always all on one side. For example, a particularly thoughtful and conscientious moral encroachment practitioner may be unsure whether there is actually anything wrong with assuming that someone is a retail employee (falsely or otherwise). After all, there is nothing wrong with *being* a retail employee. This job is socially under-valued but it is not actually disvaluable. Indeed, the moral encroachment practitioner may reasonably worry that there is instead something morally untoward about being *reluctant* to assume that someone is a retail employee, since this reluctance reinforces the pernicious ideas that working in retail is in some way shameful and that it is worse to be an employee than to be a customer. Given that there is actually nothing wrong with being a retail employee, the moral encroachment practitioner may worry that her reluctance to assume that someone occupies this social role embodies a pejorative attitude toward all actual retail employees, thereby wronging *them*, regardless of whether it wrongs the particular person whose employment status is in question. If this is what the moral encroachment practitioner in **Sports Store** thinks, then it is no longer easy to see what she should do. She cannot simply refrain from adopting "the" morally risky/costly doxastic state, since none of her doxastic options is clearly risk- and cost-free. Cases like this are cases of *moral encroachment under moral uncertainty*.

The following is a clearer example of such a case, on which we will focus for the remainder of this paper:[7]

> **Title IX Allegation.** Your student discloses to you that another student on campus has sexually assaulted her. During the standardized training that you recently received about your institution's Title IX protocols, you learned that of 100 sexual assault cases reported in your state's colleges and universities last year, the accused was disciplined in 70 of them. What should be your attitude toward the proposition that the accused in this case assaulted the student who has come to your office?

This is another case in which no doxastic option is clearly morally safe. If you are a conscientious moral encroachment practitioner then you will likely feel torn. You may be confident that significant moral costs attach to taking someone to have committed sexual assault when in fact they did not do so, especially if your doxastic state is based partly on statistical evidence. So, you may accept a moral norm that favors epistemic caution in this case: some version of the Benefit of the Doubt Norm. However, you may also be confident that significant moral costs attach to disbelieving a sexual assault allegation when the allegation is true, especially if the allegation is made to you directly as

[6] Although Fritz and Jackson mention this proposal, they do not ultimately endorse it, and they describe it as "heterodox" (p. 9). By contrast, Johnson King and Babic take their position to follow from Bayesian orthodoxy.

[7] We find this case clearer because the competing moral norms are a little bit more intuitive and specifying some statistical evidence (which is important for our model) is a little bit more natural. But the Sports Store case can be a case of moral encroachment under moral uncertainty just as well, as we explained above. Indeed, any case where we can identify competing moral norms bearing on our doxastic states, whose relative strength is to some extent uncertain to us, is such a case.

someone whom the testifier believes she can trust. So, you may also accept a moral norm that favors epistemic boldness in this case: some version of the Victim Deference Norm. (Here we use the terms "caution" and "boldness" without any assumptions as to which is preferable; we remain neutral on the first-order moral question that is the subject of your uncertainty in our example, in part because we ourselves are uncertain about it).

Many people feel the pull of both of these types of moral norm. But few of us are fully convinced of a specific view about their precise relative moral significance. For instance, few would claim that the Victim Deference Norm is precisely eleven times as important as the Benefit of the Doubt Norm. It is not even clear what sort of evidence could rationally convince us of something so precise. Moreover, many people feel the pull of both of these types of moral norm without having a clear idea of exactly what each of them amounts to, nor of exactly what sort of doxastic behavior each one calls for in a case like **Title IX Allegation**. All of this is typical of everyday moral life; ordinary moral reasoning does not consist in the smooth application of exceptionless general principles, but rather in a messy and intricate attempt to identify the morally significant aspects of our circumstances as exhaustively as we can, determine the relative importance of each of these considerations and the ways in which they interact, and thereby determine what we all-things-considered ought to do. This is rarely done with certainty.

In a case like **Title IX Allegation**, then, many of us who feel the pull of both the Benefit of the Doubt Norm and the Victim Deference Norm would be at least somewhat unsure about how to weigh these competing moral norms against one another. But this means that, even if we sincerely want to be good moral encroachment practitioners, we may be unsure how to practice moral encroachment in cases where we are unable to first resolve our underlying moral uncertainty.

A few more words about the **Title IX Allegation** example are in order before we continue. Some might be inclined to view this example as a classic case of *normative conflict*, in which the requirements of epistemic rationality and those of morality pull in different directions. But we do not think that this is correct. That's because it is unclear what either epistemic rationality or morality requires in this kind of case, and thus unclear whether they pull in different directions. On the moral side, as we have observed, some considerations favor epistemic caution – those to which the Benefit of the Doubt Norm calls our attention – while other considerations favor epistemic boldness – to which the Victim Deference Norm appeals. The relative importance of these norms is unclear, and so it is unclear what morality requires of you in this case. Meanwhile, on the epistemic side, it is unclear what the doxastic impact of the statistical evidence that you received during your training should be. The idea that this sort of evidence rationalizes a high credence in the guilt of the accused is controversial (Nelkin, 2000; Colyvan et al., 2001). Accordingly, one might think that your student's testimony is the only really weighty evidence in this case, since the reference class information may seem moot – and its impact on your credence highly irresilient – in the face of the student's testimony. And the strength of the testimonial evidence will depend on a lot of details that we have left unspecified; whether the testimony is uncontested, whether it is corroborated by your student's peers, and so forth. Thus the requirements of epistemic rationality in this case are likewise unclear.[8]

Title IX Allegation remains a case of moral encroachment under moral uncertainty regardless of how we fill out the details of your evidence in the case, and thus regardless of the strength of your epistemic position with respect to the proposition that the accused student is guilty of sexual assault. Indeed, the case would be a case of moral encroachment under moral uncertainty even if no statistical evidence were involved and your only evidence came from students' (perhaps

[8] Thanks to an anonymous referee for encouraging us to consider these points.

conflicting) testimony – though here we are more interested in cases that involve some interplay between statistical evidence and testimony, especially in the context of worries about noisy data (Section 4.3). In general, no matter what the nature and strength of your evidence in any given case may be, you may wish to respond to this evidence in a manner that is sensitive to the moral risks and/or costs at hand – you may wish, in other words, to be a moral encroachment practitioner. And you are in a case of moral encroachment under moral uncertainty just as long as you would like to be a moral encroachment practitioner, but you feel the pull both of at least one moral consideration that favors epistemic caution (like the Benefit of the Doubt Norm) and of at least one moral consideration that favors epistemic boldness (like the Victim Deference Norm), and you are unsure of these competing norms' precise relative moral importance. This is the kind of predicament – the predicament of a moral encroachment practitioner facing moral uncertainty – that we want to think about in the present paper.

3. Prevailing approaches and their limits

Extant approaches to moral encroachment have little to say about moral uncertainty. As such, they offer little in the way of helpful advice to the moral encroachment practitioner in a case like **Title IX Allegation**. This section provides an overview of their limits in such cases.

On some views, moral encroachment works just like pragmatic encroachment: moral stakes change the amount or type of evidence that is needed for someone's belief to be justified or to constitute knowledge. These views are often motivated using pairwise case comparisons, in which we feel fine about someone's doxastic state in a "low stakes" case – where nothing especially bad will happen if the person's belief turns out to be false or their credence highly inaccurate – but we have intuitive reservations about their doxastic state in a "high stakes" case that is otherwise a minimal pair. (See Fritz (2017) for a defense of this view and illustrative examples, and see Bolinger (2020a) for a similar view, also motivated partly by pairwise case comparisons, cast in credal terms.) Proponents of this sort of view can accept the traditional evidentialist tenet that one's belief is justified iff it is adequately supported by one's evidence. But they suggest that high moral stakes alter what is required for a body of evidence to be adequate, either by requiring *more* evidence in total, or by requiring evidence of a certain *sort* (e.g. non-statistical evidence). The message is that a body of evidence that would suffice for justified belief – and perhaps knowledge – in a low-stakes context no longer suffices when the moral stakes are high. And the upshot is that you should suspend judgment in the high-stakes cases until you acquire more or better evidence.

Views in this camp face two problems in accommodating cases like **Title IX Allegation**. Both problems stem from the fact that this is a case in which there are high moral stakes on both sides: believing that the accused committed sexual assault when in fact they did not is highly morally costly (as the Benefit of the Doubt Norm suggests), and disbelieving or suspending judgment about this proposition when it is in fact true is also highly morally costly (as the Victim Deference Norm suggests). The first problem is that this means that it is unclear what impact the moral stakes have *overall* on the amount and/or variety of evidence that suffices for justification. To put the point crudely, we might say that it is unclear whether the moral stakes are "pushing the epistemic standards up" or "pulling them down" overall. As a result, it is unclear whether the usual injuction for high-stakes cases – to suspend judgment until you acquire more or better evidence – kicks in or not, as it is unclear whether the moral stakes have altered what it takes for a body of evidence to be adequate in a way that is overall favorable or disfavorable to your current evidence.

The second problem is that in cases like **Title IX Allegation**, although there is nothing wrong with gathering more evidence, it is not remotely clear that suspending judgment is the appropriate doxastic state to adopt while one does so. In this case, unlike in traditional encroachment cases, the moral encroachment practitioner cannot regard suspending judgment until she has acquired more evidence as a "safe" or "neutral" option. That is because the costs involved in doubting survivors of sexual assault are costs that we incur whenever we *don't believe* their allegations – whether we actively disbelieve them or we merely suspend for the time being. What matters to victims of sexual assault is that their testimony is presumptively taken as true. But, of course, what matters to those accused of sexual assault is precisely the opposite of this: that they themselves are presumed innocent until proven guilty. Suspending judgment, then, is not neutral. Indeed, there is no neutral ground in cases like this. But this means that the injunction to avoid high-stakes doxastic states is of little use in **Title IX Allegation**, since such an injunction rules out *all* of the doxastic options. In this case, no doxastic attitude is morally risk-free.[9]

These problems would be lessened if the moral encroachment practitioner had reason to think that the moral risks on either side were equal, or roughly equal, in magnitude. In **Title IX Allegation**, if you had reason to think that the Benefit of the Doubt Norm and the Victim Deference Norm were of equal moral importance – such that presuming someone is guilty of sexual assault when they are in fact innocent is equally as bad as failing to believe a survivor – then you could continue to regard suspending judgment and gathering more evidence as the appropriate response to the moral stakes. (Notice, though, that this would not be because you regard suspension as neutral and risk-free; it would instead be because you regard suspension as the uniquely correct response to your precise view about the moral risks – namely, that they are symmetric.) But that is not so. Instead, the relative magnitude of the moral costs of the two types of epistemic error in this case is unclear. That is precisely what makes it a case of moral uncertainty. Thus, cases of moral encroachment under moral uncertainty are particularly difficult for views that enjoin us to mind moral risks by sticking to morally "safe" or "neutral" doxastic options while we inquire further. In such cases morally safe options do not exist, and nor is it clear what it would take for us to have enough evidence to conclude inquiry and take a doxastic risk.

Similar problems arise for views that focus on the moral costs involved in certain sorts of epistemic behavior *regardless* of the truth or accuracy of the agent's (resultant) doxastic states. On some views, forming beliefs with certain contents – especially pejorative contents about people to whom you bear a morally significant relationship – on the basis of certain bodies of evidence is a way of wronging the individuals that your beliefs are about. You doxastically wrong your ex-alcoholic spouse by believing that they have fallen off the wagon based on the wine stains on their sleeve (Basu and Schroeder, 2019). You doxastically wrong your customer by believing that they will leave you or your colleagues a low tip based on the fact that they belong to a racial group that frequently tips low (Basu, 2019a). And so on.[10] On a different sort of view in this camp (Marušić and White, 2018), we can doxastically wrong people by failing to believe *them* when they attempt to tell us something – that is to say, when we treat them as a source of information like any other, updating on their testimony just to the degree that our evidence indicates them to be reliable rather than having the kind of disposition to accept what they say at face value that is characteristic of trust. This view holds that our epistemic interlocutors are entitled to the more trusting approach and that failing to take it doxastically wrongs

[9] There are also versions of this case in which you do not have time to gather additional evidence and must take a stance – for example, if you are the university's Title IX administrator and are at the end of the disciplinary process. And there are versions of the case in which suspending is in practice tantamount to disbelieving – for example, if your doing anything other than actively believing your student will result in her deciding not to contact the university's Title IX office. In these versions of the case, too, suspending judgment is not the epistemic panacea that it is sometimes made out to be.

[10] See also Fritz and Jackson (MS) for an argument that the considerations supporting these views about full beliefs extend to analogous views about high credence.

them by adopting Strawson (1962)'s "objective stance" toward them, failing to treat them as an agent. Proponents of either view in this camp can then argue that a doxastic state that wrongs somebody is for that very reason epistemically unjustified. Alternatively, they can concede that such states may be justified but argue that they are nonetheless states that we ought not to adopt in light of the moral costs.[11]

It should be clear what the problem is. In a case like **Title IX Allegation**, just as a stakes-based encroachment view might end up having to say that all of your doxastic options carry high moral risks, so too might a doxastic wronging view entail that all of your doxastic options would wrong somebody. You would doxastically wrong the accused by believing that they are guilty of sexual assault based in part on the statistical evidence that you received during your recent training – or any other statistical evidence indicating that incidences of women lying about sexual assault are few and far between. And you would doxastically wrong your student by failing to believe *her*, especially given the nature of your relationship – as her professor, you are *in loco parentis* morally even if not legally – and the deeply personal nature of her choice to disclose to you. If a doxastic state that wrongs someone is for that very reason unjustified, then in a case like this there may be no justifiable doxastic option. Likewise, if such states (though perhaps justified) are impermissible in light of their moral costliness, then in a case like this there may be no doxastic option that remains permissible. Again, then, it is hard to see what recommendation this sort of view can offer to the moral encroachment practitioner in a case like **Title IX Allegation**. In such complex cases, again, the view appears to rule everything out.

Another view – the last that we will discuss – that is grouped under the "moral encroachment" label is the view that moral considerations bear on whether doxastic states constitute knowledge via some general modal conditions on knowledge. Many epistemologists think that someone can know a proposition P only if her evidence rules out all of a certain set of error possibilities that are alternatives to P – perhaps the *relevant* alternatives, or the *salient* alternatives, or those that obtain at *close worlds*, or those such that it would not be *abnormal* for them to obtain. So far, this is not a moral encroachment view. But it can be made into one. Moss (2018b) argues that moral principles can direct our attention toward certain error possibilities, which will then be salient and, if we cannot rule out these possibilities, will then prevent our doxastic states from constituting knowledge. Slightly differently, Moss (2018a) suggests that the moral costs that would follow if an error possibility obtained can suffice to make this possibility one that an agent must rule out in order for her doxastic states to constitute knowledge, whether or not this possibility is in fact salient to the agent.[12] This view is like a stakes-based view and unlike a doxastic wronging view in that it is the moral costs of *error* that undermine a doxastic state's epistemic status, rather than the moral costs involved in simply adopting or not adopting a certain doxastic state. But the mechanism by which the moral costs of error undermine knowledge on this view is different from that of a stakes-based view – they render error possibilities salient or relevant, rather than raising the bar at which evidence suffices for justification – which reflects differences in the two views' underlying pictures of the nature of epistemic justification and of knowledge.[13]

[11] It is controversial whether that final "ought" can be understood as an *epistemic* ought, giving us a case in which moral standards encroach upon the epistemic, or must be understood as a *moral* ought, giving us a direct clash between the verdicts of epistemic rationality and those of morality. We take no stand on the best way to spell out the view. The problem we are about to describe arises either way.

[12] Moss focuses on credences, but analogous positions can be applied to full beliefs; the view would be that beliefs do not amount to knowledge if the subject is unable to rule out the alternatives in the relevant set, and that moral considerations affect what is in that set in either or both of the two ways that Moss proposes.

[13] See Bolinger (2020b) for an overview of the moral encroachment literature that classifies positions not only according to whether it is moral risks or costs on which they focus but also according to the mechanism by which they take these risks/costs to impact our doxastic states' epistemic status.

Unlike the other views surveyed in this section, Moss (2018b) does issue a clear recommendation in cases like **Title IX Allegation**. For Moss proposes a specific moral rule directing our attention toward error possibilities: the "rule of consideration", as she calls it, states that we should keep in mind the possibility that individuals can be exceptions to accurate statistical generalizations. In **Title IX Allegation**, then, we should keep in mind the possibility that the accused might be an exception to the accurate statistical generalization that most people accused of sexual assault on local college campuses are guilty. Since our evidence does not rule out this error possibility, Moss's view implies that a credal state according to which the accused is probably guilty would not constitute knowledge. If the moral encroachment practitioner wants to avoid credal states that fail to constitute knowledge in light of the moral risks, it is clear what she should do: she should think that the accused is not probably guilty. While this is a clear recommendation, it is not a compelling one, given that there are also moral costs associated with failing to believe your student in this case. Moss's picture effectively offers a way to implement the Benefit of the Doubt Norm while remaining silent about the Victim Deference Norm. So this view says something clear about our case of moral encroachment under moral uncertainty only by effectively ignoring half of the source of the uncertainty.

This may not be true of Moss (2018a), depending on how the details of her position are fleshed out. Here Moss does not say anything specific about which moral considerations can make an error possibility into a relevant one, one that an agent must rule out in order for her doxastic states to constitute knowledge. So it is consistent with her view that the moral badness of mistakenly taking the accused to be guilty makes this error possibility relevant (incorporating the Benefit of the Doubt Norm) *and* the moral badness of mistakenly taking the accused to be innocent *also* makes this other error possibility relevant (incorporating the Victim Deference Norm). In this case, Moss's view is in a similar position to those that we have surveyed so far. If one must be able to rule out all morally costly error possibilities in order for one's doxastic state to constitute knowledge, then in a case like **Title IX Allegation** there is no doxastic state that can constitute knowledge, since no doxastic state can meet this epistemic burden. It is therefore hard to see what sort of recommendation the view could issue in a case of moral encroachment under moral uncertainty; once again, the view appears to rule everything out.

To be sure, we are not arguing that it is *impossible* for any of the preceding views to provide guidance about moral encroachment under moral uncertainty. These views could be developed or extended in ways that speak to such cases. Our point is that this development has not yet occurred. Furthermore, for these views to issue helpful recommendations to the moral encroachment practitioner under conditions of moral uncertainty, they will have to be spelled out in much more detail than they have been thus far. The simple injunction to avoid doxastic states that carry high moral risks or costs is inadequate when there is no risk-free or neutral doxastic position, as in **Title IX Allegation**.

More strongly, we suspect that all encroachment views cast in terms of full beliefs, rather than credences, will be unable to issue plausible recommendations in cases of moral encroachment under moral uncertainty. In most such cases, the competing considerations are too fine-grained to be captured adequately with only three doxastic states. For example, in **Title IX Allegation** you are faced with the prospect of combining testimony, statistical evidence, and background knowledge together with your assessment of the relative moral standing of the Victim Deference Norm and the Benefit of the Doubt Norm. Given the number of open parameters in this problem, it is unlikely that we can lump everyone who might face such a situation into one of just three bins: those who should believe the victim, those who should disbelieve the victim, and those who should suspend judgment. Rather, what is appropriate in the moral encroachment practitioner's circumstances will likely depend on precisely what she thinks about the veracity of her sources as well as precisely how she settles the moral trade-off. It is our interest in capturing this complexity that drives us to develop a model enabling the agent to practice moral encroachment under moral uncertainty in a way that is sensitive to these multi-level fine-grained judgments.

Notice also that the moral encroachment practitioner's uncertainty concerns hypotheses about the relative *degree* of importance of the multifaceted moral considerations at stake in her circumstances. She may entertain the hypothesis that a certain norm is exactly twice as important as another alongside the hypotheses that the former is three times as important, or only 0.5 times as important, and so on. And we can expect these hypotheses to be quite fine-grained; for instance, few bodies of evidence would support the hypothesis that one norm is exactly twice as important as another without also supporting the hypothesis that the former is 2.001 times as important, 2.0001 times as important, and so on. Since the moral encroachment practitioner faces uncertainty concerning hypotheses about moral considerations' relative degrees of importance, it seems sensible to accommodate its impact on her doxastic states using an approach that allows doxastic states to also come in degrees.

4. The pyramid of uncertainty

4.1 Starting point

There is one view on moral encroachment that provides a promising starting point for accommodating moral uncertainty. Johnson King and Babic (2020)'s Bayesian encroachment view focuses on how moral encroachment practitioners can balance the competing risks of increasing credence in a proposition (such as the accused's guilt in **Title IX Allegation**) that turns out to be false and of decreasing credence in a proposition that turns out to be true. This risk-balancing approach, which Johnson King and Babic use to identify agents' priors and their responsiveness to evidence allows, in principle, that there may be moral costs on both sides and that they need not be equal in magnitude. This is just the sort of assessment that must be made in cases of moral encroachment under moral uncertainty. It will therefore be our starting point.

Johnson King and Babic emphasize that a standard Bayesian view about updating one's credences on new evidence, properly understood, *is* an encroachment view. This is because changes to an agent's credences are evaluated, on this approach, relative to their expected inaccuracy. And every way of evaluating accuracy or inaccuracy corresponds to some way of striking a balance between the badness of graded false positive error (i.e. increasing one's credence in a falsehood) and graded false negative error (i.e. decreasing one's credence in a truth). More precisely, every *strictly proper scoring rule* – the way of evaluating the accuracy of credences that Bayesian epistemologists ordinarily employ – can be derived from the agent's underlying *epistemic risk function* – a function that encodes her assessment of the relative badness of these two types of graded mistakes – through a pair of ordinary differential equations.[14]

The agent's epistemic risk function can also be used to select an appropriate prior. This is accomplished by identifying the shape of the agent's epistemic risk function and determining how much risk she is willing to assume, where the least risky prior is the one such that the agent's inaccuracy score if the proposition in question turns out to be true is equal to her inaccuracy score if the proposition turns out to be false. On this view, then, the agent's assessment of the relative badness of the two types of epistemic error determines what her initial doxastic state should be, and, as a result, also affects her sensitivity in responding to later evidence.

Johnson King and Babic further emphasize that there is no neutral

[14] See Babic (2019) for a detailed development of the epistemic risk framework.

position on the costs of epistemic error. There is no Archimedean point from which to assess accuracy absent of all value judgments. Some formal epistemologists hold that a uniform prior (i.e. one that assigns equal probability to every possible value of a random variable) represents indifference or a total lack of information. But Johnson King and Babic observe that this is in fact the prior that someone would have if she were committed to the view that false positive error and false negative error are *exactly equally bad* under an ordinary symmetric scoring rule. To say that graded false positive and negative error are exactly equally bad is obviously not a way of avoiding making a value judgment; rather, it is itself a value judgment. And it is as particular and precise a value judgment as any other. On this approach, then, there is simply no way to avoid taking a stand on the relative magnitude of the different error costs.

Further, Johnson King and Babic explain that a standard Bayesian picture does not dictate the basis for one's assessment of the magnitude of these costs. Bayesian orthodoxy simply requires that agents have *some* attitudes toward the two types of epistemic error (explicit or otherwise), since it is impossible to evaluate the accuracy of credences without them. But there is nothing in the ordinary Bayesian framework that precludes assessing the costs of error in pragmatic or moral terms. Why the costs are what they are is something that Bayesianism itself cannot answer; the agent has to bring these judgments with her into an inference problem. Indeed, pragmatic and/or moral assessments of the costs of error are widespread in practice – as, for instance, when a weather forecaster's hurricane predictions are made in a way that hedges against false negative mistakes, on the grounds that predicting a hurricane when there is no hurricane (false positive) would be mildly inconvenient whereas failing to predict an actual hurricane (false negative) would be a disaster. Hence their claim is that a standard Bayesian view simply *is* a moral encroachment view.[15]

Johnson King and Babic's approach still falls short, however, in cases like **Title IX Allegation**. While their approach allows in principle for the possibility that moral costs attach to both types of epistemic error, the model they develop involves a one-sided case analogous to **Sports Store**. They are concerned to show that a moral encroachment practitioner can have a weighted prior that makes her correspondingly less sensitive to statistical evidence suggesting that certain individuals are more likely than others to possess socially disvalued traits based on their membership in groups in which those traits are prevalent. To illustrate this point, they simply pick an epistemic risk function that roughly corresponds to an assessment of the moral costs of error that seems intuitively compelling in their case. They then examine how an individual with this stipulated risk function would update on new evidence. Moreover, and more problematically for present purposes, Johnson King and Babic assume that every moral encroachment practitioner's attitude toward epistemic risk is fully specified, reflecting a particular and precise assessment of the relative costs of the two types of epistemic error. In other words, they assume that the agent is certain of a particular epistemic risk function *ex ante*. As a result, its shape is an exogenous input to the inference problem and how the agent came to be certain of it is not the subject of their work.

But this is precisely what cannot be assumed in cases of moral encroachment under moral uncertainty. That is because cases of moral uncertainty are cases in which the moral encroachment practitioner's

[15] Of course it does not *have* to be the moral costs of error on which the agent focuses. To repeat: Bayesianism itself does not say anything about which costs to focus on. The agent could instead insist that her attitudes to error come from purely alethic considerations or purely pragmatic considerations. Or she may insist that they are determined by consulting the last two digits of her phone number or of her zip code. Bayesianism cannot dictate how she comes up with these costs (i.e. it cannot dictate the shape of her utility function, telling her what matters to her).

epistemic risk function is *not* fully specified; the agent is unsure about the relative badness of the two types of epistemic error. For example, she may be unsure about the relative moral significance of the Benefit of the Doubt Norm and the Victim Deference Norm. When the agent is uncertain about the shape of her epistemic risk function, and in turn her scoring rule, we cannot use Johnson King and Babic's method to determine her prior.

So, while this model has the potential to issue clear recommendations to moral encroachment practitioners in cases like **Title IX Allegation**, it still needs to be developed in a way that makes room for moral uncertainty. We undertake this development in the remainder of this section. In the final subsection, we will then draw on an existing extension of Johnson King and Babic's model to explore what happens when a morally uncertain moral encroachment practitioner obtains imperfect or misleading evidence, which in this project we precisify as data that are *themselves* uncertain or "noisy". This is a ubiquitous situation, which would typically obtain in instances like our central case, **Title IX Allegation**; it is likely in situations like this that the available statistical data are a somewhat imperfect representation of the true state of affairs (i.e. some students who committed sexual assault were not disciplined and/or some who did not were disciplined). This is an aspect of the inference problem that should not be ignored. Our final model combines uncertainty about noisy data with moral uncertainty to yield some striking and informative results.

4.2 *Our model*

Since the material to follow gets somewhat technical, we start by providing an overview of our strategy. First, we will summarize the mathematical details of Johnson King and Babic's approach, which can handle moral encroachment without moral uncertainty. That is to say, we will explain how an agent's epistemic risk function encodes her attitudes to graded error and how this function determines the shape of her scoring rule and, in turn, her prior.

Next, we will develop our extension of this model to handle cases of moral uncertainty. Our key move is to create room for the possibility that the agent is unsure about the relative badness of the two types of epistemic error, and therefore about her own epistemic risk function and her own priors. A helpful way to describe this might be to say that the agent has *higher-order uncertainty* about what her prior should be, due to her moral uncertainty about the stakes in the inference problem.

Finally, we will explain how our model can handle not only higher-order uncertainty but also uncertainty about the representativeness of one's data. In other words, it can be easily expanded to capture situations where an agent must update on noisy statistical evidence.

Let's begin. Johnson King and Babic focus on what they call *pernicious predictive inference*. In cases of pernicious predictive inference, we obtain some data that bear on the underlying proportion, θ, of members of a population (e.g. college students facing Title IX proceedings, young people of color in sports stores) who possess some undesirable and/or socially undervalued trait (e.g. having committed sexual assault, being a sports store employee). More precisely, in a group of n people, let t represent the number of people who possess such a trait and $n-t$ the number of people who do not. Our observations are therefore coming from the following likelihood function[16]:

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

[16] That is to say: this is the function describing, in terms of θ, n, and t, the "random draws" (observations) from the relevant population. To make this terminology easier to understand, consider that we could describe the tosses of a coin with bias 0.6 as being drawn from a likelihood function given by $0.6^t(1-0.6)^{n-t}$, in which case the probability of observing $t=3$ heads in a series of $n=3$ three tosses would be equal to $0.6^3(1-0.6)^0 = 0.216$.

We must then use the available data to estimate the probability that a new individual from the population possesses the sensitive trait. Johnson King and Babic's idea is that this sort of predictive inference can be "pernicious" because false positive error involves falsely imputing a sensitive trait to someone who does not in fact possess it.

As Johnson King and Babic emphasize (following Huttegger (2017) and Lindley and Phillips (1976), among others), we can describe the Bayesian approach to making predictive inferences in cases like this in terms of a three-step procedure. First, we need a prior probability distribution for the value of θ, the proportion of individuals in the population who in fact possess the trait. This is the object of uncertainty in standard Bayesian epistemology – it is the central unknown and unobservable (or only imperfectly observable) quantity, which the agent must estimate and use as a basis for predictions about individuals. If you are totally ignorant about θ, then your distribution might be uniform over the entire interval $[0, 1]$. But this is rarely reasonable. However, there is a flexible prior in cases like this that allows us to model a variety of information states about θ. It can be written as follows:

$$\pi(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{2}$$

where $\alpha > 0$ and $\beta > 0$. This is a sensible prior to use in the case of inference on a proportion, because it can take just about any desired shape depending on the values we assign to α and β. Indeed, if $\alpha = \beta = 1$ then the entire expression on the right hand side is equal to 1 and it becomes the uniform prior. Further, the mean of this distribution is given by $\alpha/(\alpha + \beta)$. Notice that this prior looks similar in form to the likelihood (1), except that we have introduced two new parameters, α and β. They will be very important in what follows, as they are how we model the agent's attitudes toward the badness of false positive and false negative mistakes.

The second step of the process is to update your prior distribution on the available data. For instance, if you start with a uniform prior and then observe 10 people from the target population, all of whom possess the relevant trait, then your updated distribution about θ – your *posterior distribution* – should shift toward 1. This step is straightforward as we use Bayes' Rule to update the prior on what was observed. Given our setup above, the posterior distribution can be written as follows:

$$\pi(\theta|t, n, \alpha, \beta) \propto \theta^{\alpha+t-1}(1-\theta)^{\beta+n-t-1}. \tag{3}$$

Notice that the posterior distribution (3) is of the same form as the prior distribution (2), except that we have a new α, given by the sum of our old α and the number of observations possessing the trait, t. And we have a new β, given by the sum of our old β and the number of observations not possessing the trait, $n - t$.

The third step of the process is also straightforward and generally mathematically determined.[17] Suppose we want to make a prediction about whether the next person we observe will possess the relevant trait. Say that we make predictions on $\widetilde{X}$, where $\widetilde{X} = 1$ represents possessing the trait and $\widetilde{X} = 0$ represents not possessing it. To accomplish this, we should use the predictive distribution for $\widetilde{X}$, given by:

$$\begin{aligned}\Pr(\widetilde{X} = 1|t, n, \alpha, \beta) &= \int_0^1 \Pr(\widetilde{X} = 1|\theta)\pi(\theta|t, n, \alpha, \beta)d\theta \\ &= \mathrm{E}[\theta|t, n, \alpha, \beta] \\ &= \frac{\alpha + t}{\alpha + \beta + n}.\end{aligned} \tag{4}$$

[17] This third step can be sensitive to the choice of utility function.

In other words, we should base our prediction on the posterior mean. Huttegger (2017) refers to this expression as the Generalized Rule of Succession and shows that it follows from several modest assumptions about the structure of our data, which will be satisfied in the types of cases we consider. (From a decision-theoretic perspective, the posterior mean minimizes expected square error loss.) Cases like **Sports Store** and **Title IX Allegation** fit this pattern, as they are also cases of predictive inference based (at least partly) on data generated by (1).

Because the second and third steps of this procedure are for the most part mathematically determined, the flexibility exists substantially in the first step: the selection of one's prior. This is where moral encroachment enters the picture. Johnson King and Babic emphasize that the norms of epistemic rationality on a standard Bayesian picture do not single out a uniquely rational prior. These norms are permissive; they instruct us only to ensure that our credence functions obey the axioms of probability and to update them by Bayes' Rule. Agents' priors are instead determined by their estimates of the relative badness of false positive and false negative graded error. The agent selects priors that minimize epistemic risk as they see it.

Here is how this works. Let $s(p, I_A)$ be the agent's scoring rule for the probability assigned to an arbitrary proposition A, $p = Pr(A)$, where $I_A = 1$ if A is true and 0 otherwise. For example, A may represent the proposition that the accused student in **Title IX Allegation** is guilty. Figure (1) below depicts two possible scoring rules for such a scenario. The prior distribution for θ is $\pi(\theta)$, as in (2). A *symmetric* scoring rule is indifferent between approaching inaccuracy in the false positive direction and in the false negative direction (left panel) whereas an *asymmetric* scoring rule takes one of these two types of error to be worse than the other (right panel). Johnson King and Babic require that $\mathrm{E}[\theta] = p^*$ where p^* satisfies $s(p,1) = s(p,0)$. This implies that the mean of θ will be where the agent's inaccuracy is the same whether or not the proposition is true or false (see Figure 1 for a visual depiction of this requirement). But recall, as we noted on p. 35, that the mean of θ can be expressed in terms of α and β: in particular, $\mathrm{E}[\theta] = \alpha/(\alpha + \beta)$. Thus, their framework imposes a requirement on the permissible values of α and β, through the location of p^*, which is determined by the agent's attitudes toward epistemic risk with respect to A.[18] So, if you are more worried about false positive mistakes with respect to some trait, then α must go down. And if you are more worried about false negative mistakes, then β must go down.

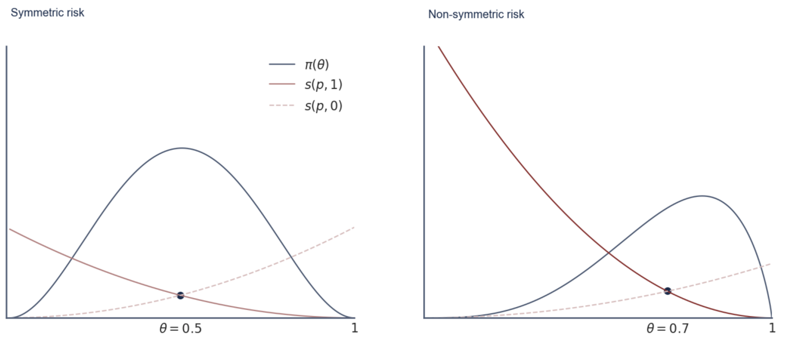

Figure 1: This figure depicts two pairs of possible scoring rules (red curves), and a permissible prior associated with each pair (blue curve). The distribution for θ is $\pi(\theta)$. A symmetric scoring rule is indifferent between approaching inaccuracy in the false positive direction and in the false negative direction (left panel) whereas an asymmetric scoring rule is not (right panel).

[18] In this project, we describe p^* in terms of where $s(p,1)$ intersects with $s(p,0)$ – i.e. the point where there is no accuracy uncertainty, hence the risk-free point. More generally, however, p^* is the minimum of the formal epistemic risk function, as articulated in Babic (2019). We do not go into the details of the epistemic risk function here as they are not necessary. It is enough if the reader sees that p^* corresponds to the least risky point in the sense that where it obtains there is no uncertainty about one's accuracy outcomes – they will get the same score whether A is true or false.

There are two important points about this setup. First, the constraint that $\mathrm{E}[\theta] = p^*$ requires only that the agent adopt a credence whose inaccuracy score is the same if the proposition in question is true as it is if the proposition in question is false. But this requirement underdetermines the overall shape of the prior, since many different distributions can share the same mean. How we narrow the choice set down, for Johnson King and Babic, depends on how conservative we ultimately want to be in responding to new evidence: a prior distribution that is densely peaked around the mean (i.e. one where $(\alpha + \beta)$ is large) will shift more slowly in response to new evidence, whereas one that is very diffuse (i.e. one where $(\alpha + \beta)$ is small) will be more responsive to evidence. Thus the moral encroachment practitioner should select a particular prior based on their assessment of the stakes in their decision problem, the relative costs of mistakes, and the anticipated quality of their evidence.

Second, and more importantly for our purposes here, in cases like **Title IX Allegation** you may be unsure as to which of the two scoring rules from Figure (1) – or many, many others – accurately reflect the relative moral badness of the two types of graded epistemic error. This is due to your moral uncertainty: you do not know whether the Victim Deference Norm or the Benefit of the Doubt Norm is more important and to what extent. Indeed, this is the hallmark of problems like **Title IX Allegation**. In practice, this means that you will not know which values you should assign to α and β.

This creates a problem for Johnson King and Babic's model. Observed "positives" (t) and "negatives" ($n - t$) from the data are supposed to be added to the agent's initial values of α and β (respectively) to determine the overall shape of the posterior distribution. The quantities α and β are the parts of the model that reflect the agent's opinion about the relative badness of the two types of graded epistemic error. But morally uncertain agents do not have a settled opinion on this. And so we cannot stipulate the values of α and β in advance, as Johnson King and Babic do: under moral uncertainty, the agent is not in a position to specify α and β. As a result, the morally uncertain agent's attitude toward epistemic risk is not precisely determined. Our morally uncertain agents will then be unsure where their risk-free point p^* is – i.e. the point that determines their set of permissible priors. Such agents are, in effect, uncertain of their scoring rule. To our knowledge, this is the first paper to examine such a situation.

This takes us outside the scope of Johnson King and Babic's model. But there is a natural way to extend their approach to cases involving moral uncertainty: **we can add an additional dimension to the model reflecting the agent's uncertainty about α and β.** We thus allow α and β themselves to be unknown quantities. The agent will have probability distributions over the true value of these quantities – which is equivalent to a distribution over the hypotheses about the relative importance of the moral considerations (e.g. the Benefit of the Doubt Norm and the Victim Deference Norm) relevant to assessing the costs of graded false positives and false negatives.[19]

The idea that agents might have probability distributions over α and β themselves is fairly intuitive if we think about it informally for a moment. Morally uncertain agents are usually not completely clueless as to the relative badness of the two types of epistemic error. In **Title IX Allegation**, for instance, while you are unsure of the precise relative importance of the Benefit of the Doubt Norm and the Victim Deference Norm, you may well have a rough sense of which is more important. And you will presumably be willing to decisively rule out some outlandish hypotheses, such as the hypothesis that the former norm is a billion times more important than the latter. (If so, then you

[19] This is an equivalent way of representing moral uncertainty to modeling the agent as having a mixture of prior distributions, whose weights correspond to the relative standing of the two competing norms. But whereas the mixture approach would become cumbersome as the uncertainty increases, our approach can handle just about any situation.

are not *maximally morally spineless* – a notion that we discuss further in Section 5.)

A flexible parametric form that can be used to represent a wide range of states of uncertainty about α and β is a normal distribution truncated at 0. Substantively, all this means is that we want a flexible shape like that offered by the normal distribution, but because α and β cannot take negative values (the competing moral norms cannot have *negative* importance!), we will truncate the distribution at 0. This is not necessarily the only, or even best, distribution to pick for α and β, but we will use it to illustrate our model. So, let

$$Z(t|\mu,\sigma) = \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right]. \tag{5}$$

Then, the prior for α and β may be given by:

$$\pi(\alpha|\mu_\alpha,\sigma_\alpha) = \frac{Z(\alpha|\mu_\alpha,\sigma_\alpha)}{\int_0^\infty Z(t)dt}, \tag{6}$$

and,

$$\pi(\beta|\mu_\beta,\sigma_\beta) = \frac{Z(\beta|\mu_\beta,\sigma_\beta)}{\int_0^\infty Z(t)dt}. \tag{7}$$

Notice that we now have two further pairs of meta-meta-parameters, μ and σ, where each pair governs the prior distribution for one of α and β. The full joint prior distribution can then be written as:

$$\pi(\theta,\alpha,\beta|\mu_\alpha,\sigma_\alpha,\mu_\beta,\sigma_\beta) = \pi(\theta|\alpha,\beta)\pi(\alpha|\mu_\alpha,\sigma_\alpha)\pi(\beta|\mu_\beta,\sigma_\beta). \tag{8}$$

Visually, we can represent the model in terms of a directed Bayesian graph as follows.

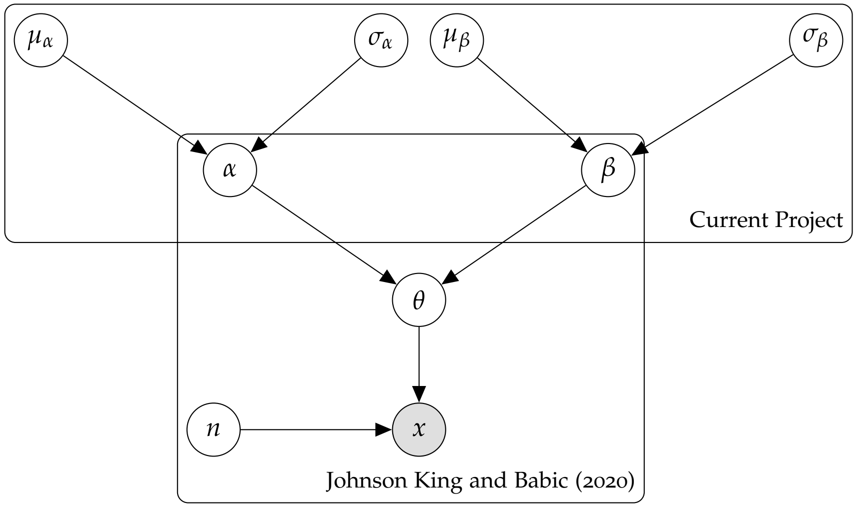

Figure 2: A model for representing higher-order uncertainty. Johnson King and Babic develop the portion in the lower box, whereas we generalize the approach by adding the remainder. In the graph, θ is the agent's estimate of the true population proportion, which is influenced by her assessment of the costs (including moral costs) of the two types of graded epistemic error, encoded by α and β. In our model, the agent's assessments of these costs do not take precise values. Instead, the agent is uncertain about these values too, and this moral uncertainty is determined by μ and σ.

One nice feature of this model is that it is almost arbitrarily scaleable. If the agent were to become uncertain about the values of μ and σ, we would add further distributions over the values of these quantities – i.e. a further layer in the upper box of Figure (2) – and so on up. The procedure for going from nth-order uncertainty to $(n+1)$th-order uncertainty is the same as the procedure for going from 1st order uncertainty to 2nd order uncertainty. The literature on moral uncertainty has begun to grapple with the issue of *higher-order uncertainty* – that is, uncertainty as to the correct approach to cases involving moral uncertainty – and our approach is well poised to capture it, since our model can easily represent uncertainty as far up as the agent's doxastic state permits.[20]

Despite the added complexity, this approach still allows for updating via Bayes' Rule and still yields precise, determinate posterior estimates as a result. Our uncertain agent will observe some data, such as those provided in **Title IX Allegation**. She will then update the full model on those data. After updating, her posterior distribution will be proportional to the joint distribution of the priors and the likelihood, as follows:

$$\pi(\theta,\alpha,\beta|x,\mu_\alpha,\sigma_\alpha,\mu_\beta,\sigma_\beta) \propto f(x|\theta)\pi(\theta|\alpha,\beta)\pi(\alpha|\mu_\alpha,\sigma_\alpha)\pi(\beta|\mu_\beta,\sigma_\beta). \tag{9}$$

The difference, as compared to Johnson King and Babic (2020), is just that we have scaled the random quantities up a level to reflect the agent's uncertainty about the relative importance of the applicable moral norms.

However, one important respect in which our model differs from that of Johnson King and Babic is that it is now difficult to compute algebraically, due to the added complexity. We can no longer exploit the similarity between the prior and the likelihood – as we did with equations (1) and (2) – in order to update by simply adding numbers of observed instances to the values of α and β. For a more complicated model like this, the solution is to approximate the posterior distribution using a Markov Chain Monte Carlo algorithm. We have relegated most of the details of this strategy to the Appendix, but the basic idea is that instead of trying to mathematically derive a complex posterior distribution, we approximate it through a computational random sampling procedure. In particular, we use an algorithm known as Hamiltonian Monte Carlo (Duane et al., 1987; Neal, 1995) implemented in a general purpose Bayesian probabilistic programming language called Stan (Carpenter et al., 2017) (we provide the associated code in the Appendix).

For illustration, suppose that the prior distributions for α and β are as follows:

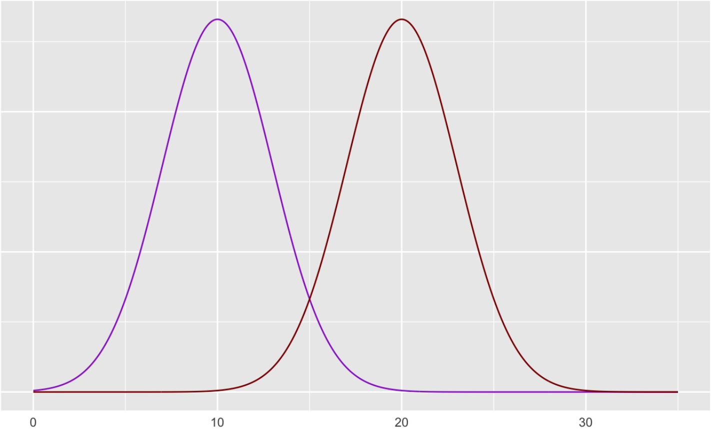

Figure 3: Prior distributions for α (purple) and β (red).

Here α and β follow a normal distribution, to the right of 0, with a mean of 10 and 20, respectively. This is a specification of the general model we have developed above. What it means, in **Title IX Allegation**, is that you are fairly confident that falsely disciplining an innocent person is considerably worse than failing to believe victims and allowing the guilty to go unpunished – as, we suspect, many college professors

[20] Note that moral uncertainty and non-moral uncertainty are not mutually exclusive. Sometimes there are evidential or physical reasons to structure μ and σ a certain way, while at other times the reasons come from considerations of epistemic risk, and at still other times they come from both sources – for example, if we obtain some evidence about which error type is more likely.

are.[21] The standard deviation for both distributions is set to 3 for the sake of illustration. The prior for α admits a large range of values; you are .95 confident that the true value of α is in (4, 16). Meanwhile, for β you are .95 confident that the true value is in (14, 26). As we can see, though, there is some overlap between the distributions. This means that you are not completely certain that the Benefit of the Doubt Norm is more important than the Victim Deference Norm. This feature of the model reflects the central aspect of cases like **Title IX Allegation**: one must practice moral encroachment while being uncertain about the direction from which and the speed at which the moral encroaches.

When we update this model on the data in **Title IX Allegation**, the posterior mean of your estimate of the true population proportion becomes 0.64. That is a bit lower than the sample mean – the proportion of students investigated and found to have committed sexual assault – which, we earlier stipulated, is 0.7.[22] This difference reflects the fact that you were antecedently confident that falsely punishing an innocent person is considerably worse than failing to believe victims and allowing the guilty to go unpunished. Your asymmetric attitude to the moral costs of graded error has thus encroached upon the updating process.

Notice the difference between this model and a model that does not have any higher-order uncertainty (such as Johnson King and Babic's model). If you were certain that the values of α and β were simply equal to, say, the mean values of their respective prior distributions as we have stated them – 10 and 20 – then the posterior mean could be computed algebraically, and it would be 0.61. That is a little bit lower than in our model. It is also a little bit further away from the sample mean of 0.7 than we get when we add moral uncertainty into the picture. This is to be expected: when your priors are more "firm", so to speak, they exert a greater pull on the data. By comparison, if in our model we set the standard deviations for the prior distributions over both α and β to 10 (rather than 3 as above – i.e. your prior distributions are *very* diffuse) then the posterior mean would be 0.68, which is almost equal to the sample mean. In this case, you have so much uncertainty about the relative costs of the two types of epistemic error that your estimate of these costs plays almost no role in your subsequent prediction. This is also to be expected: the more unsure you are about which of the Benefit of the Doubt Norm and the Victim Deference Norm is more valuable and to what extent, the less weight your prior estimates of these values can exert on the data. This brings your posterior estimate closer to the sample mean. We summarize the above estimates in Table 1 and discuss these points further in Section 5.

Sample mean	0.7
Posterior mean with significant moral uncertainty	0.68
Posterior mean with modest moral uncertainty	0.64
Posterior mean without moral uncertainty	0.61

Table 1: Summary of posterior estimates of the accused's guilt in **Title IX Allegation** depending on different assumptions about the degree of moral uncertainty with respect to the Benefit of the Doubt Norm and the Victim Deference Norm.

4.3 *Noisy data*

So far we have assumed that the moral encroachment practitioner takes her data at face value. In **Title IX Allegation**, this implies that the 70 out of 100 students who were disciplined are all and only those who were in fact guilty of sexual assault. Such uncritical trust in one's data is

[21] Of course, there is nothing special about this particular epistemic risk profile – we are certainly not recommending it. Indeed, we will be somewhat critical of it in the next section. Here we are just using it to illustrate our central idea.

[22] For the sake of this example, "found guilty" is equivalent to "guilty", as we clarify in Section 4.3, where we consider cases in which this may not be the case.

not entirely unusual – for instance, if one hears that 7 out of a group of 10 puppies are Pomeranians, then one usually takes this proportion for granted rather than worrying about the risk of non-Pomeranian puppies masquerading as Pomeranians or vice versa. But this uncritical attitude can become problematic as the risk of misclassification increases. In our hypothetical sample of 100 students, it is near certain that some of them would have been falsely classified – either disciplined when they were in fact not guilty, or acquitted (or otherwise relieved of all complaints[23]) when in fact they were guilty. Thus, while the observed proportion of guilty to innocent students in **Title IX Allegation** is 70-30, the true (but unknown) proportion might be 71-29, 24-76, 90-10, and so forth. This raises the question: How should a moral encroachment practitioner update her beliefs when the evidence includes some misclassification?

One solution would be to follow Jeffrey (1983, 1992)'s approach for updating on uncertain evidence, where the posterior belief is a mixture of the posterior beliefs conditioning on the various possible observations weighted by their probabilities. But notice that in **Title IX Allegation**, where the sufficient statistic is the sum of guilty defendants (i.e. no additional information from the sample would bear on inference about the proportion), there are 101 various possible observations. Thus, Jeffrey conditioning would require assigning each of them a probability, and then computing that large sum. While this is already exceedingly difficult with 100 observations, it would be near impossible with 100,000 or 10 million.

As a result, we take a different approach, which can be easily couched within the general model we have developed above. Following Babic et al. (2021); Gaba (1993); Gaba and Winkler (1992) and Winkler and Gaba (1990), we can simply further generalize our model to capture situations with noisy data. To do so, we add additional unknown quantities – additional parameters, which will be estimated in the model – corresponding to the misclassification rate(s).

Recall that θ, in our model above, represents the true-but-unknown proportion of individuals who possess the relevant trait. To capture misclassification, we need to add room in the model for the fact that there is some proportion of disciplined people who are not actually guilty – let η represent this proportion. Likewise, there is some proportion of not-disciplined people who are actually guilty – let λ represent this proportion. Now, instead of θ representing the true-but-unknown proportion of individuals who are in fact guilty, this proportion is given by $\theta(1-\eta) + (1-\theta)\lambda$. Let τ represent this value. Then, instead of $1-\theta$ representing the true-but-unknown proportion of individuals who are not guilty, this proportion is given by $1-\tau = (1-\theta)(1-\lambda) + \theta\eta$. These equations are each the sum of two terms, because there are two ways to fall under the relevant grouping: to belong to the category (i.e. guilty or not guilty) and to be correctly classified, or to not belong to the category but be misclassified as belonging to it.

Hence, the likelihood function can be written as,

$$\ell(x|\theta,\lambda,\eta) \propto \left[\theta(1-\lambda) + (1-\theta)\eta\right]^{x}\left[1 - \left[\theta(1-\lambda) + (1-\theta)\eta\right]\right]^{n-x}. \tag{10}$$

τ is then the new unknown quantity relevant to our predictive inferences with noisy data (i.e. the analogue to θ). As before, we will need a prior distribution for τ, which will be updated on new observations and used for prediction and estimation. And in order to specify a prior for τ, we now need to specify priors for η and λ as well as α, β, μ and σ. We will assume that the priors for η and λ also follow a truncated normal distribution, just as the priors for α and β in Equations (6) and

[23] They could be relieved of all complaints without being acquitted – for instance, if the accuser is persuaded to drop the case before the investigations are complete.

(7):

$$\pi(\eta|\mu_\eta, \sigma_\eta) = \frac{Z(\eta|\mu_\eta, \sigma_\eta)}{\int_0^\infty Z(t)dt}, \tag{11}$$

and,

$$\pi(\lambda|\mu_\lambda, \sigma_\lambda) = \frac{Z(\lambda|\mu_\lambda, \sigma_\lambda)}{\int_0^\infty Z(t)dt}. \tag{12}$$

This is again a flexible and plausible assumption for these new parameters, as they too must take values between 0 and 1 – i.e. the misclassification rate lies somewhere between no mistakes and all mistakes. The full joint prior distribution can then be written as the product of the priors, as in Equation (8), and the posterior distribution is proportional to the product of the full prior and the likelihood, as in Equation (9). Since the prior is now specified in terms of eight parameters, a μ and σ for each of α, β, η, and λ, we omit the full expressions for the joint prior and posterior. But the ensuing model can be readily visualized as follows (and the code is provided in the Appendix):

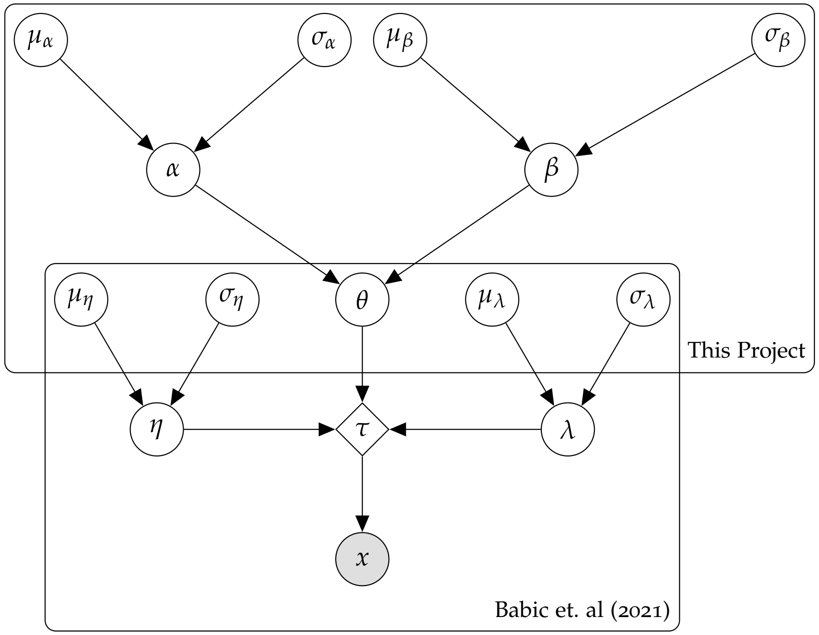

Figure 4: A model for representing higher order (moral) uncertainty when data are noisy. We have omitted n (the number of observations) for simplicity, which is fixed, as before.

For illustrative purposes, we will suppose that $\mu_\eta = 0.1$, $\mu_\lambda = 0.3$, $\sigma_\eta = \sigma_\lambda = 0.01$. This means that your priors encode a non-trivial degree of confidence that there were some mistakes in how the individuals in the sample were classified. Perhaps, for instance, you heard during your training that your institution's Title IX proceedings have been criticized as procedurally inadequate. We will keep everything else as specified in our previous illustrative example without noise. That is, $\mu_\alpha = 10$, $\mu_\beta = 20$, $\sigma_\alpha = \sigma_\beta = 3$. This is now the model we update on the **Title IX Allegation** data.

After updating on the data in **Title IX Allegation**, the mean of your posterior distribution for τ will be 0.55: a *lot* less than the posterior estimate of 0.64 in the noise-free case, and also a good deal less than the sample mean of 0.7. The value of the data is thereby diminished, and your prior plays a relatively larger role in the final estimate.

This is the opposite effect of what we described in the previous subsection. When moral uncertainty is introduced into the prior estimate of the true population proportion, the posterior estimate is pulled toward the data. But when uncertainty about the extent to which one's data are representative of the underlying reality is introduced, the posterior estimate is pulled (back) toward the prior. Your final estimate of the probability that a new individual from the relevant group possesses the relevant trait will depend on how you settle these two sources of uncertainty. These results are worth pausing to reflect on, as we do in the next and final section.

5. Upshots

In **Title IX Allegation**, if your attitudes are as we have stipulated them in our final illustrative example (with both moral uncertainty and noisy data), then your credence that the accused student did indeed sexually assault your student after updating on the data will be 0.55 – i.e. your posterior estimate for τ. That may seem highly objectionable, since your credence is substantially lower than the base rate itself. Nonetheless, this is how your credences will develop if your attitudes are as we have stipulated them: if you are morally uncertain, but you suspect that the moral facts are more concerned with protecting defendants than with believing victims, and you are convinced that your data includes some misclassification, but you suspect that the proportion of disciplined students who are innocent is larger than the proportion of non-disciplined students who are guilty. Were we to make the reverse stipulations – that believing victims is more important than protecting defendants, and that the data include proportionately more students who commit sexual assault and get away with it than students who are erroneously disciplined – then we would see a pull in the opposite direction, such that you are particularly inclined to believe your student. These observations have two striking practical implications.

The first implication pertains to a character that Johnson King and Babic characterize as "not necessarily irrational, but just a jerk" (Johnson King and Babic, 2020, p. 98). This is someone who faces a case like **Sports Store** with perfectly symmetric attitudes toward epistemic risk. The jerk does not think that either type of epistemic error is worse than the other – notwithstanding the fact that one of these errors involves falsely assuming that someone bears a socially disvalued trait on the basis of their race, which seems quite bad (hence the "jerk" label for someone who is indifferent to this badness). Such a character will have the uniform prior that formal epistemologists often take to represent indifference, which exerts minimal weight on the posterior. When the jerk obtains some information about the observed frequency of a trait within a group, their estimate of the probability that the next group member that they encounter possesses the trait will then be approximately equal to the observed frequency (e.g. approximately 0.7 if the frequency is 70%). Johnson King and Babic argue that the problem with this person is not that they violate any formal requirement of epistemic rationality, but simply that they adopt a morally bad attitude: they are indifferent between making a racist mistake and a non-racist mistake, which is morally bad, since one morally ought to be averse to racism. The jerk, then, does not need to justify their epistemic rationality. Their real uphill battle would be in giving a moral argument in support of this dubious attitude toward epistemic risk.

The first striking implication of our model is that someone who is maximally morally uncertain – who is morally *spineless*, as we called it earlier – will often behave, epistemically, just like a jerk. The morally spineless person is someone who is unwilling to rule out *any* hypotheses about the relative moral significance of the competing considerations at stake. They are not even willing to rule out, say, the outlandish hypothesis that the Benefit of the Doubt Norm is a billion times more important than the Victim Deference Norm. Their doxastic attitude is one of total uncertainty over all putative combinations of moral facts. This moral uncertainty is extremely drastic, making the agent's priors over moral facts so diffuse that their attitudes toward epistemic risk ultimately carry very little weight in the ensuing inference problem, leaving almost nothing but the observed frequency as the basis of their posterior distribution. As a result, their associated predictions will be nearly indistinguishable from those of someone who is completely certain that the moral costs of error are exactly equal – that is to say, of the jerk. In general, the more morally spineless one is, the more one's updating behaviour will resemble that of the jerk. Meanwhile, someone who is maximally morally certain will be pulled by her normative attitudes in a fixed direction. As a result, her prior will exert more weight on the data than that of a morally uncertain agent, and much more than a morally spineless agent.

This first implication should be interesting for theorists of moral uncertainty and traditional Bayesians as well as for moral encroachment theorists. For the first of these groups, our result echoes a sentiment sometimes expressed in the literature on moral uncertainty by those with externalist leanings (especially Weatherson (2019), pp. 43-44): complex and difficult cases call for *moral bravery*, for taking a stand, perhaps in spite of one's ambiguous evidence. Now, we have not vindicated this sentiment with anything like the level of generality at which it is sometimes advanced. We certainly do not think that *any* amount of moral uncertainty is cowardly and that morality demands complete certainty of a particular, precise set of first-order moral facts. On the contrary, we think that having some degree of moral uncertainty is clearly reasonable in response to the messy evidence that moral agents often face, although we have assumed rather than arguing for this point in the present paper. Nonetheless, we do vindicate a certain kind of exasperation that one might feel when faced with with the morally spineless. For one might quite naturally feel exasperated by the spineless person's failure to *appreciate the value* of that to which they cannot commit. And our model vindicates this sentiment with the striking result that maximal moral uncertainty looks, in practice, just like committed indifference.[24]

To be sure, this provides only a *prima facie* case against moral spinelessness. Those who are already skeptical of the moral relevance of moral uncertainty might take our striking result (alongside the assumption that jerky behavior is morally bad) as a point in favor of their view. But we have not argued for the moral irrelevance of moral uncertainty; on the contrary, our working assumptions are that it is entirely reasonable for normal human agents to harbor a moderate amount of moral uncertainty and that this uncertainty is relevant to how we ought to behave epistemically, rendering some degree of epistemic "hedging" appropriate.[25] Our model is effectively a formal model of one way in which someone might hedge. To respond to our *prima facie* case against complete moral spinelessness, philosophers sympathetic to the idea that moral uncertainty is itself morally relevant have several options. They could show that their view does not condone complete spinelessness, but rather something more moderate. Or they could provide some countervailing considerations in favor of condoning spinelessness despite our *prima facie* case against it; for instance, they could identify further differences between the spineless person and the jerk that rationalize more favorable attitudes toward the former than the latter, or they could reject moral encroachment and thereby maintain that spinelessness does not have the epistemic implications that we have described.

For the second of the above groups (traditional Bayesians), the

[24] To be a little bit more specific, our view is that maximal moral uncertainty constitutes exasperating spinelessness whereas minimal moral uncertainty – i.e. complete confidence in a single precise moral theory – constitutes unwarranted dogmatism. Moral bravery lies in-between, although the borderlines are vague and subject to debate.

[25] For more on the idea of "hedging" in response to moral uncertainty see, for example, Nissan-Rozen (2015) and Hicks (2019).

interest of our first implication is in its highlighting that diffuse priors are not always as epistemically virtuous as they are widely taken to be. In Bayesian inference, it is important to be sufficiently open-minded so as to not rule out possible observations *a priori*. For instance, one would not want to assume *ex ante* that a certain coin is not two-sided.[26] However, what our model suggests is that while epistemic humility in the Bayesian sense (of diffuse priors) may be a virtue for ordinary empirical uncertainty, it can turn to vice – the vice of spinelessness – when it comes to moral uncertainty. And this seems entirely reasonable. Even if you are considerably uncertain about the Benefit of the Doubt Norm's and the Victim Deference Norm's relative importance, surely you can rule out *some* hypotheses, such as the 1 billion : 1 ratio. That would not be a sensible way to deal with accusations of sexual assault. And this reasonable ruling-out is precisely what will stop you from acting like a jerk. It pays, then, to be at least minimally resolute – to be somewhat willing to take a stand.

Our second striking implication builds on the first. It is that the effect of moral uncertainty just described – the more morally uncertain you are, the more you end up acting like a jerk – can be counterbalanced by suspicions about the misleadingness of your evidence. Moral uncertainty lessens the impact of one's prior credences on one's posterior credences, as we have observed (Table 1). But there is also something that correspondingly lessens the impact of one's data on one's posterior credences: suspicion that the data are noisy and therefore unrepresentative. Uncertainty about the reliability of the data can thus provide a counterweight to the impact of moral uncertainty on posterior credences, tipping the scales in the other direction. And, again, this makes intuitive sense: if someone is radically morally uncertain but trusts her data, then she will be heavily swayed by the data – the one thing she is sure of – whereas someone who is relatively confident in her assessment of the moral risks will be swayed by the data to a lesser degree, and someone who is confident in her assessment of the moral risks but significantly mistrustful of her data will barely be moved by them at all. This means that it is possible for someone to become increasingly morally uncertain without acting increasingly like a jerk, provided that their moral uncertainty is accompanied by comparable uncertainty about the extent to which their data are a faithful reflection of the underlying facts.

These two striking implications have an important practical upshot for the last of the groups enumerated above – i.e. moral encroachment practitioners. This upshot is of particular interest to those of us who would like not only to practice moral encroachment ourselves, but also to encourage others to form and revise their doxastic states in a manner that we deem morally laudable – being slow to assume that the person in front of them in **Sports Store** is an employee, for instance. The upshot is that there are two ways to get people to do this. First, you can reduce their moral uncertainty: argue for the badness of assuming that someone works in retail based simply on their social demographic, so as to convince them that the relative costs of the two types of epistemic error could not assume any ratio under the sun; rather, they must instead fall within a moderate range circumscribed by at least tolerably defensible attitudes to epistemic risk. Second, you can increase their uncertainty about the reliability of their data: give them reasons to think that reported observed frequencies do not match the true proportions of individuals in various groups who possess various traits, or do not match the (purported) "propensity" of individual group members to possess the traits.[27] The more noisy they take their data to be, the less weight these data will exert on their posterior distributions

[26] In statistics, this virtue often goes by the name Cromwell's Rule (Lindley, 1991), after a story about Oliver Cromwell. Following his role in the execution of Charles I during the Second English Civil War, Cromwell wrote a letter to the Church of Scotland urging against the appointment of Charles's son as Scotland's King. In that letter, he writes: "I beseech you, in the bowels of Christ, think it possible that you may be mistaken" (Carlyle, 1845).

[27] We use scare quotes here because we are doubtful that it even makes sense to speak of the "propensity" of, say, a young person of color to become a shop assistant or a college student to commit sexual assault.

and, consequently, their predictive inferences. Similarly, the less morally uncertain they are, the more weight their prior – encoding their assessment of moral risk – will exert on their posterior, and, consequently, their predictive inferences.

If we want people to update in a morally laudable manner, then these are two quite different strategies that we can pursue in order to get them to do it: examine the data with a critical eye, or reduce their uncertainty with moral or political argument. Neither bears directly on the underlying Bayesian machinery, but both strategies are of critical practical importance.

Appendix

In this appendix we produce the computational details supporting Section 4. We use a Markov Chain Monte Carlo algorithm known as Hamiltonian Monte Carlo (HMC) to approximate the posterior distributions of our models (Duane et al., 1987; Neal, 1995). For the case without noise (Section 4.2) we implement the following model:

$$\begin{aligned} X &\sim \text{Binomial}(n,\theta) \\ \theta &\sim \text{Beta}(\alpha,\beta) \\ \alpha &\sim \text{Truncated Normal}(10,3) \\ \beta &\sim \text{Truncated Normal}(20,3) \end{aligned}$$

To do so, we use the following Stan code, a general purpose Bayesian programming language (Carpenter et al., 2017):

```
MME1.stan:
data {
 int < lower = 1> n;
 int < lower = 1, upper = n> Y;
}
parameters {
 real < lower = 0, upper = 1 > theta;
 real < lower = 0 > alpha;
 real < lower = 0 > beta;
}
model {
    Y ~ binomial(n, theta);
    theta ~ beta(alpha, beta);
    alpha ~ normal(10, 3) T[0, ];
    beta ~ normal(20, 3) T[0, ];
}

model_path <- "MME1.stan"
model_mme = stan_model(model_path)
stan_data <- list(Y = 70, n = 100)
fit_main <- sampling(model_mme, data = stan_data,
            warmup = 10000, iter = 100000, chains = 2,
            cores = 1,
            thin = 1,
            control =
            list(adapt_delta = 0.99, stepsize = 0.001,
            metric = "dense_e"))
```

For the case with noise (Section 4.3), we implement the following model:

$$\begin{aligned} X &\sim \text{Bin}(n,\tau) \\ \tau &= \theta(1-\lambda)+(1-\theta)\eta \\ \theta &\sim \text{Beta}(\alpha,\beta) \\ \alpha &\sim \text{Truncated Normal}(10,3) \\ \beta &\sim \text{Truncated Normal}(20,3) \\ \eta &\sim \text{Truncated Normal}(0.1,0.01) \\ \lambda &\sim \text{Truncated Normal}(0.3,0.01) \end{aligned}$$

This model is again fitted to the **Title IX Allegation** data in Stan with the following change to the file above:

```
MME2.stan:
data {
 int < lower = 1> n;
 int < lower = 1, upper = n> Y;
}

parameters {
 real < lower = 0, upper = 1 > lambda;
 real < lower = 0, upper = 1 > eta;
 real < lower = 0, upper = 1 > theta;
 real < lower = 0 > alpha;
 real < lower = 0 > beta;
}

transformed parameters {
    real < lower = 0, upper = 1 > tau;
    tau = theta*(1-lambda)+(1-theta)*eta;
}

model {
    Y ~ binomial(n, tau);
    lambda ~ normal(0.3, 0.01) T[0, 1];
    eta ~ normal(0.1, 0.01) T[0, 1];
    theta ~ beta(alpha, beta);
    alpha ~ normal(10, 3) T[0, ];
    beta ~ normal(20, 3) T[0, ];
}
```

This completes the Appendix. Note that our goal is to identify a reasonable model for illustrating the interaction between ordinary uncertainty, moral uncertainty, and data noise, as described in Section 5. There are many different choices we could make in articulating the setup. But those illustrative choices are not central to our argument. The key is to highlight the conceptual framework we develop for reasoning about moral encroachment under moral uncertainty.

Acknowledgments

We are both very grateful to two anonymous referees and to our editor Brian Weatherson for extremely helpful feedback and discussion. Thanks also to Jamie Fritz, Georgi Gardiner, Ilia Tsetlin, and Ralph Wedgwood for helpful conversation and formative feedback on early versions of this material.

Boris Babic would additionally like to thank the Social Sciences and Humanities Research Council of Canada, and the Schwartz Reisman Institute for Technology and Society at the University of Toronto, for their financial support.

References

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

Babic, Boris, Anil Gaba, Ilia Tsetlin, and Robert L. Winkler (2021). Normativity, Epistemic Rationality, and Noisy Statistical Evidence. *British Journal for the Philosophy of Science* (Accepted April 30, 2021).

Basu, Rima (2019a). The Wrongs of Racist Beliefs. *Philosophical Studies 9*(176), 2497–2515.

Basu, Rima (2019b). What We Epistemically Owe To Each Other. *Philosophical Studies 4*(176), 915–931.

Basu, Rima and Mark Schroeder (2019). Doxastic Wrongings. In Brian Kim and Matthew McGrath (Eds.), *Pragmatic Encroachment in Epistemology*, pp. 181–205.

Bolinger, Renee (2020a). The Rational Impermissibility of Accepting (Some) Racial Generalizations. *Synthese 6*(197), 2415–2431.

Bolinger, Renee (2020b). Varieties of Moral Encroachment. *Philosophical Perspectives 1*(34), 5–26.

Carlyle, Thomas (1845). *Oliver Cromwell's Letters and Speeches*. New York: Scribner.

Carpenter, B., A Gelman, Lee M. Hoffman, B. Goodrich, M. Betancourt, M.A. Brubaker, J. Guo, P. Li, and A.R. Stan (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software 76*(1), 1–32.

Colyvan, M., H.M. Regan, and S. Ferson (2001). Is it a Crime to Belong to a Reference Class? *Journal of Political Philsophy 9*(2), 168–181.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid Monte Carlo. *Physics Letters B 195*(2), 216–222.

Fritz, Jamie (2017). Pragmatic Encroachment and Moral Encroachment. *Pacific Philosophical Quarterly 98*(S1), 643–661.

Fritz, Jamie (2020). Moral Encroachment and Reasons of the Wrong Kind. *Philosophical Studies 177*(10), 3051–3070.

Fritz, Jamie and Liz Jackson (MS). Belief, Credence, and Moral Encroachment. *Synthese* (Forthcoming).

Gaba, Anil (1993). Inferences with an Unknown Noise Level in a Bernoulli Process. *Management Science 39*(10), 1179–1197.

Gaba, Anil and Robert L. Winkler (1992). Implications of Errors in Survey Data: A Bayesian Model. *Management Science 38*(7), 913–925.

Gardiner, Georgi (2018). Evidentialism and Moral Encroachment. In K. McKain (Ed.), *Believing in Accordance with the Evidence: New Essays on Evidentialism*.

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

Hicks, A. (2019). Moral Hedging and Responding to Reasons. *Pacific Philosophical Quarterly 100*(3), 765–789.

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

Jeffrey, Richard (1983). *The Logic of Decision* (2nd ed.). Chicago: University of Chicago Press.

Jeffrey, Richard (1992). *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

Johnson King, Zoë and Boris Babic (2020). Moral Obligation and Epistemic Risk. In Mark Timmons (Ed.), *Oxford Studies in Normative Ethics*, Volume 10, pp. 81–105.

Lindley, Dennis (1991). *Making Decisions*. New York: Wiley.

Lindley, Dennis V. and L.D. Phillips (1976). Inference for a Bernoulli Process (A Bayesian View). *The American Statistician 30*(3), 112–119.

Marušić, Berislav and Stephen White (2018). How Can Beliefs Wrong?: A Strawsonian Epistemology. *Philosophical Topics 46*(1), 91–114.

Moss, Sarah (2018a). Moral Encroachment. *Proceedings of the Aristotelian Society 2*(118), 117–205.

Moss, Sarah (2018b). *Probabilistic Knowledge*. Oxford: Oxford University Press.

Neal, R. M. (1995). An Improved Acceptance Procedure for the Hybrid Monte Carlo Algorithm. *Journal of Computational Physics 111*(1), 194–203.

Nelkin, Dana (2000). The Lottery Paradox, Knowledge, and Rationality. *Philosophical Review 109*(3), 373–409.

Nissan-Rozen, I. (2015). Against Moral Hedging. *Economics and Philoso-*

phy 3(1), 1–21.

Strawson, P.F. (1962). Freedom and Resentment. In Gary Watson (Ed.), *Proceedings of the British Academy*, Volume 48, pp. 1–25.

Weatherson, Brian (2019). *Normative Externalism*. Oxford: Oxford University Press.

Winkler, Robert L. and Anil Gaba (1990). Inference with Imperfect Sampling from a Bernoulli Process. In Nicholas Longford et. al. (Ed.), *Bayesian and Likelihood Methods in Statistics and Econometrics*, pp. 303–317. North-Holland.

Worsnip, Alex (2021). Can Pragmatists Be Moderate? *Philosophy and Phenomenological Research 102*(3), 531–558.