Comments on Gendler's, "The Epistemic Costs of Implicit Bias"

Andy Egan, Rutgers University and Arché Philosophical Research Centre

Draft of July 24, 2011

I'm grateful for the opportunity to comment on Tamar Gendler's extremely interesting and thought-provoking paper. I've profited a great deal from reading and thinking about it, and I haven't found much to disagree with. I'll quibble at a few points, but mostly what I'll do here is say why some initially attractive ways of defusing some of Gendler's pessimistic conclusions don't actually look very promising, and draw some connections with discussions of conflict between epistemic and non-epistemic norms in other domains.

1. Alief

Gendler frames much of her discussion of the negative consequences of living in a society structured by objectionable racial categories in terms of her notion of *alief*, and chalks up much of the epistemic costliness of living in such a society to the fact that we are the sorts of agents that have alief-based (or anyway, alief-including) psychologies. This makes it tempting to object to her diagnosis of our compromised epistemic position by objecting to her alief-including psychology, and argue that we're not in the bad epistemic situation Gendler takes us to be in, because we don't have the kind of psychology that she thinks we have. This is, I think, a blind alley. Skepticism about alief ought not give rise to skepticism about the substantive points of the paper. (I'm skeptical about alief, but pretty much convinced about everything else she says here.) In this section, I'll say a bit more about Gendler's psychological category of *alief*, present some grounds for skepticism about the invocation of alief, and say why I think that this sort of reasonable alief-skepticism doesn't undermine anything that Gendler is doing here.

In her paper in this volume, Gendler characterizes alief as follows: "to have an *alief* is to have an innate or habitual propensity to respond to an apparent stimulus in a particular way. In paradigm cases and on strict usage, this response involves an automatized representational-affective-behavioral triad." (Gendler, this volume, PAGE REF – p 8 of 32 in the word document I'm looking at.)

The notion of *alief* is one that plays a prominent role in much of Gendler's work (Gendler, 2007; 2008a; 2008b), and so it's probably worth saying a bit more about what she's got in mind. This will help to get clear on why although the invocation of alief is here – as elsewhere -- potentially contentious, the features of alief that make for the contentiousness aren't the ones that are doing the heavy lifting in generating the epistemic problems that Gendler is concerned with in this paper. Elsewhere (Gendler, 2008b), she gives a more detailed presentation of the idea:

> To have an alief is, to a reasonable approximation, to have an innate or habitual propensity to respond to an apparent stimulus in a particular way. It is to be in a mental state that is (in a sense to be specified) *a*ssociative, *a*utomatic and *a*rational. As a class, aliefs are states that we share with non-human *a*nimals; they

> are developmentally and conceptually *a*ntecedent to other cognitive attitudes that the creature may go on to develop. Typically, they are also *a*ffect-laden and *a*ction-generating." ((Gendler, 2008b)

She continues:

> Traditional propositional and objectual attitudes are two-place affairs. A subject believes (that) b or desires (that) d or hopes (that) h or fears (that) f. But alief, as I propose to use the term, involves a relation between a subject and an entire associative repertoire, one that paradigmatically includes not only representational (or 'registered') content, but also affective states, behavioral propensities, patterns of attentiveness, and the like. There is no natural way of articulating this, but—as a reasonable (if cumbersome) approximation—we can say that a subject in paradigmatic state of alief is in a mental state whose content is representational, affective and behavioral: she alieves r, a, b. Though this usage is approximate—and in that sense, misleading—it helps to emphasize the ways in which thinking in terms of alief differs from thinking in terms of the traditional cognitive and conative attitudes.
>
> Examples will make this usage clearer. Consider again the frog going for the BB, the puppy batting at the mirror, and the suspended man trembling in the cage. In each of these cases the norm-discordant behavioral tendency can be explained by an alief with content that might be expressed, among other ways, as follows. The frog alieves (all at once, in a single alief): small round black object up ahead; appealing in a foody sort of way; move tongue in its direction. The puppy alieves (again, all at once): dog-shaped dog-motiony creature in front of me; attractive and threatening in a my-size-conspecific sort of way; engage in (play-)fighting. The suspended man alieves (all at once): high up above the ground right now, dangerous scary place to be, tremble. ((Gendler, 2008b), p559)

There are two central strands to Gendler's specification of aliefs. One strand concerns their downstream effects, and the difficulty we have regulating them through straightforward cognitive means. As Gendler emphasizes repeatedly, aliefs can measurably influence subsequent behaviors and attitudes, giving rise to belief-like, affective, and motivational contents that may be in conflict with, and aren't short-circuited or cancelled out by, a subject's explicitly held, consciously endorsed beliefs, attitudes, and intentions. This is in part because aliefs are (at least relatively) cognitively impenetrable. We can't (easily) excise them by reasoning, or standard processes of belief-revision. We can't (or can't easily) short-circuit their activation, or their effects. Call this (to have a manageable, though slightly misleading, name for it) the *automaticity* strand.

The other strand of Gendler's picture is that aliefs are also supposed to be instances of a psychological state that is more primitive and more basic than belief or desire, and,that rather than "alief" being a name for a cluster of distinct and independently characterizable doxastic, affective and motivational states, aliefs are simple, unified states that have belief-like, affective, and motivational aspects. Call this the *priority* strand.

Each of these two strands can be broken down further, in ways that are important, though they will not play a substantive role in my larger argument. The automaticity strand breaks down into two substrands: first, there's the automaticity of the representational, affective, and motivational responses to some class of stimuli. And second, there's the automaticity of the formation of the dispositions. Neither the particular representational, affective and motivational responses on a given occasion, nor the formation of the dispositions to have the responses, is (fully) responsive to the agent's rational, deliberate processes of thought, revision, etc. The priority strand also breaks down into two component substrands: One of these is a *unity* strand – aliefs are simple, unified states. The other is a *fundamentality* strand – they're more basic than, for example, belief and desire. (As Gendler puts it, they're "developmentally and conceptually *a*ntecedent to other cognitive attitudes" (2008b).) (Thanks to Tyler Doggett for both of these distinctions.) Though the details of my argument won't turn on these, I think it's worth noticing the structure here.

It's reasonable, I think, to wonder whether the phenomena that we see are sufficient to motivate wheeling in the big gun of a new fundamental taxonomical category for mental states. What we see is automatic, arational and non-short-circuitable representation, affect, and action which are potentially discordant with the subject's explicitly held and endorsed beliefs, attitudes and plans. (And which, even when concordant with the subject's explicitly endorsed attitudes, aren't traceable to them in the paradigmatic way.) But we already have some independently motivated, and less radical, theoretical tools available for explaining these kinds of phenomena.

We have fragmented or compartmentalized belief (Egan, 2008; 2009; Elga, 2005; Lewis, 1982; Stalnaker, 1984) and "in-between belief" (Schwitzgebel, 2001; 2002), and similar kinds of fragmentation for non-doxastic attitudes. Compartmentalized versions of conventional propositional attitudes can be expected to display the same sorts of automaticity-strand features that aliefs are also meant to have: limited behavior-guiding role, imperfect integration with the rest of the subject's attitudes, and imperfect responsiveness to many mechanisms for revision and updating. We have non-short-circuitable Spinozan mechanisms of belief-formation (Egan, 2008; Gilbert, 1991; Gilbert, Krull, & Malone, 1990; Gilbert, Tafarodi, & Malone, 1993; Huebner, 2009; Mandelbaum, 2011), by which beliefs (or in-between-beliefs) get formed automatically, without approval from the subject's conscious methods of evaluation. And we have conventional propositional attitudes realized at a sub-personal level, in such a way that they're not responsive to conscious, personal-level reflection, updating, suppression, etc. in the paradigmatic kinds of ways (see, for example, (J. A. Bargh, Gollwitzer, Lee-Chai, Barndollar, & Trötschel, 2001; J. Bargh & Chartrand, 1999; Bortolotti, 2010; Fitzsimons & J. Bargh, 2003; Fodor, 1983; Mandelbaum, 2011; Sterelny, 2003).

The phenomena Gendler's discussing, and which she invokes alief to explain, are precisely the kinds of phenomena that these sorts of psychological stories seem to predict – indeed, they're just the kinds of phenomena that motivate people to go in for these kinds of accounts. So it's appropriate, I think, to resist the call to sign on

for a sui generis mixed state until we've been given a story about why the more familiar and less radical resources aren't up to the explanatory job.

One way to put the concern, both about the motivation for introducing aliefs and about how well we understand what's being proposed, is as a concern about two of the things Gendler says in the passage quoted above. One might be worried about just what it means to say that the representational, affective, and behavioral/motivational bits of happen "all at once, in a single alief". What is it, exactly, that distinguishes this possibility from one better described as, "all together, in a tightly clustered bunch of distinct representational, affective, and motivational states"? And once we're clear on just what the single-state claim means, we might also be worried about whether or not it's true. What are the phenomena that speak in favor of a single-state account rather than a cluster account, and which can be accommodated and explained by the first, but not the second?

I think it's also reasonable to be worried about the fact that "there's no natural way to articulate" the components of the alief and their relations to each other. It doesn't seem likely that it's just a coincidence that the affective part is systematically an affective response that's to be expected, given the representational part, plus the subject's array of more conventional mental states (beliefs, desires, etc.), and that the action part is always an action that fits with the representational content and affective response, given the subject's array of more conventional mental states. (There's a good reason, presumably, that we see a lot of "high up, scary, tremble", and not so much "high up, melancholy, tap dance".) If aliefs are just primitive, unified states, it's a little bit hard to see what kind of explanation we're going to be able to give of why we tend not to see versions of those states with representational, affective and motivational aspects that wouldn't fit together well if they were separate representational, affective and motivational states. (Or do we see them, and I'm just not thinking of the right examples?) We seem to forego the possibility of a certain kind of explanation of the connections between the representational, affective and motivational phenomena if we bundle them into a single primitive state, rather than thinking in terms of a cluster of distinct states, the connections between which might admit of some interesting explanation.[1]

[1] I think that I haven't quite put my finger squarely on the worry here. But I'm pretty confident that there *is* a worry here for somebody to put a finger on. Brendan Dill may have come closer, in his comments on an earlier draft of this paper: "The real question is whether the behavior and affect that follows from the representational component of the alief is sensitive to changes in the subject's more conventional mental states. If they are, then that favors an idea that the three are distinct, since the R component would remain unchanged while the AB components would switch in response to changes in the subject's beliefs and desires. If they are not, then that favors a unity view where the three are inseparable – the alief is activated or not as a fully interconnected RAB repertoire." (Dill, p.c.) I think that this is right, with the complication that we'll need to make sure that we're taking into account not just changes in the agent's superficial, cognitively-easily-accessible conventional mental states when we're doing this test. And to the extent that it's hard to pull apart the predictions of attributions of automatic, impenetrable, etc. versions of conventional mental states from those of attributions of aliefs, it's going to be difficult to do the relevant experiments.

I don't want to linger too long on criticism of alief, though, because (as I've indicated above) aliefs aren't really a particularly load-bearing part of the story here. They give us one way to conceptualize the kinds of effects we see in members of societies in which racism is prevalent, whereby racist stereotypes, attitudes, and patterns of behavior get a grip on the cognitive, affective and motivational lives even of members of the society who explicitly and sincerely disavow racism. But they're not the only way, and however we conceptualize them, the effects are undoubtedly real.

One way of bringing this out is to recognize that it's the automaticity strand in Gendler's notion of alief that's doing the work in this paper, and it's the priority strand that's contentious. (I'm skeptical about alief because I'm skeptical about the priority strand of Gendler's account of them.) What's important for Gendler's purposes here is just the idea that there are some doxastic, affective and motivational states that (a) cluster together in this distinctive way, and (b) are cognitively impenetrable. And that's definitely true. Whether the best account of them is in terms of a primitive, more-basic-than-belief-or-desire state of *alief* or in terms of fragmented or compartmentalized versions of more conventional sorts of doxastic, affective and motivational states, is more contentious. (I for one am not yet on board.) But that also doesn't really matter for the purposes of this paper.

Indeed, the reason why it doesn't much matter whether we go along with the postulation of an alief-based psychology in this case is exactly the reason that I'm inclined to do without alief altogether: we can generate all of the same phenomena, and all of the same problems and epistemic costs, from fragmented belief and attitudes that aren't fully responsive to cognitively driven revision and whose activation isn't short-circuitable. So we can tell a very similar story to the one

---

Tamar Gendler (p.c.) suggests certain types of bizarre phobias and attractions as potential examples – if one has (to use her examples) a phobia about chocolate rabbits, or a thing about smurfs, it's very plausible that the affective and motivational bits of the phobia, or the pathological attraction, will be bad fits (given the subject's other conventional attitudes) with the representation of objects as chocolate rabbits, or as smurfs. As Gendler puts it, some good candidates for the right kind of non-separable RAB clusters are, for example, "'Smurfs, sexy, [censored]' or 'Chocolate rabbit, SCARY, retreat' – in those cases, the R pulls one way and the AB pull another – and no changes in the subject's more conventional mental state are going to alter the AB pattern. I manage to convince you that smurfs aren't sexy. You say: "I agree completely. Smurfs are completely unsexy. I have no desire to couple with smurfs." But then I show you a smurf image. And it turns you on. Or I show you the physics of candy-making and convince you that chocolate rabbits are totally harmless. And you say: "I completely agree. Chocolate rabbits are harmless and delicious! I have an intense desire to eat a chocolate rabbit. Yummy!" And then I show you a chocolate rabbit. And you faint in fear." (Gendler, p.c.)

These seem like promising cases to me, though resolving what's going on in any given case is again going to complicated by the fact that we'll want to make sure that the affective and motivational components of the would-be alief are bad fits, not just with the subject's conscious, explicitly acknowledged and endorsed beliefs, desires, etc., but also with the subject's unconscious, sub-personal, fragmented ones. One explanation of why the candy-making lessons and the smurf-sexiness debunking aren't effective in defusing the fear or attraction is that the fear and attraction aren't grounded in aliefs. Another is that they're grounded on beliefs that, on account of their relative cognitive impenetrability, aren't responsive to these kinds of rational, evidence-based methods of updating. (I get worried at this point – as does Gendler (p.c.) – about to what extent there's more than a terminological dispute here. There's a bunch more to say about this, but here probably isn't the place to say it.)

Gendler tells, but on which we appeal (for example) to fragmentation of the right kind rather than a primitive state of alief. The things Gendler says elsewhere that distinguish alief accounts from accounts according to which there's a cluster of more conventional kinds of attitudes that are not fully integrated with the rest of the agent's psychology (so as to be cognitively impenetrable, resistant to rational revision, etc.), aren't really doing any crucial work here – the troublemaking states don't need to have any priority relative to belief, etc.

So it would be a mistake to think that the lessons of Gendler's paper hinge on a commitment to a psychology of aliefs. The alief picture is one convenient way to conceptualize the phenomenon, but nothing much changes, as far as the epistemic costs of living in a society structured by racist categories goes, if you conceptualize it another way. I'll follow Gendler in talking in terms of alief in what follows, but it's worth bearing in mind that nothing much changes if, rather than thinking in terms of alief, you'd rather think in terms of clusters of cognitively impenetrable beliefs, attitudes, etc.

2. Easons

In her discussion in section 3 of her paper, Gendler offers a tentative proposal for how to understand the etiology of the loss of knowledge and decline of performance under stereotype threat:

> Now what *may be* interesting, epistemically and more generally about the stereotype threat cases is that they appear to fall somewhere in between the reason cases and the cause cases. I repeat: this is not well-worked out and I am counting on my readers to help me in thinking this through in later papers. In the stereotype threat cases, the loss of knowledge is due at least in part to a feeling of anxiety that is induced through activations of self-referential cultural stereotype aliefs whose content you disavow. This loss of knowledge isn't just the result of something straightforwardly causal like bumping your head and getting a concussion – or even just something like being hungry, or tired, or preoccupied; nor is it the result of something straightforwardly reason-based like reading a revisionist textbook or thinking through a Brain-in-a-Vat scenario. That is, it's not due to pushes and pulls and bumps and bounces – nor is it due to beliefs and arguments and reflections and persuasions. Rather, it's the result of something that – for want of a better term – I'll provisionally call an *eason*: something that is not sufficiently well-conceptualized to call a reason, but that (in a way in between a reasony and a causy fashion) *eases* us towards a certain outlook on the world.
>
> Roughly speaking, on this picture, beliefs are to reasons as aliefs are to easons: we justify beliefs by appeal to reasons; we explain aliefs by appeal to easons. Our imagined victim of stereotype threat has an *alief* with the content "'Female' applies to me and 'female' is associated with poor math performance; (anxiously) better make sure that I'm doing these math problems correctly; double-check double-check double-check." The *eason* that she has this alief is similarly complex. It consists in some sort of complex interplay among stress, anxiety-induced self-monitoring and self-regulative emotion suppression. This case is distinct from one where our subject *believes* "perhaps the answer to the third question is not in fact 17" or "I, like most females, am bad at math." If

> asked explicitly, she would deny both, and offer reasons for each that she reflectively endorses: "Of course the sum of 1, 2, 3, 4 and 7 is 17: 1+2+3+4=10, and 10+7=17" or "Of course I'm good at math: when I'm at home after school, I love to play math games on my computer." Nor is it simply a case where she is merely *caused* out of holding that 1+2+3+4+7=17: she hasn't been hit in the head by an errant baseball, or had arithmetic-scrambling electrodes implanted in her brain, or forgotten to eat breakfast, the most important meal of the day (WebMD, 2010). Rather, her alief is the eason for her knowledge-loss. (Gendler, this volume, PAGE REF)

I agree with Gendler that there seems to be an interesting not-quite-rational, not-quite-brute-causal phenomenon happening here. (I also want to take this opportunity to signal my appreciation of her public service announcement on behalf of breakfast.) But I'm not sure that the *easons* picture is the right way to think about it. So in this section, I'd like to (also tentatively) suggest a slightly different diagnosis.

What seems off to me about the easons picture is that it doesn't seem like the most informative thing that we can say about the mechanism is that it occupies some intermediate point on a spectrum between brute-causal mechanisms and rational-reason-response mechanisms. There seems to be more structure to the case than that. It's not just a matter of locating the mechanisms as intermediate between paradigmatic rational causation by psychological mechanisms on the one end of the spectrum and paradigmatic non-rational causation by non-psychological mechanisms (booze, blows to the head, etc.) on the other. At least part of what's distinctive about the cases in question seems to be that they're cases of non-rational causation by psychological mechanisms. The response *is* brought about by "beliefs and arguments and reflections and persuasions", it's just that it's not brought about by them in the usual (or anyway, the paradigmatic) rational-response-ish kind of way. Instead, it's brought about by some bit of belief, perception, argument, thought, etc. triggering an association, which calls up the stereotype-concordant aliefs, which give rise to some stereotype-concordant thoughts, feelings, and motivations which run in parallel with the subject's consciously-endorsed thought processes and start horning in on the agent's behavior guidance and conscious thought processes. Or anyway – that seems like a diagnosis worth exploring.[2]

[2] Brendan Dill (p.c.) offers what seems to me a very promising suggestion for getting a tighter handle on the phenomenon. He suggests that we need to distinguish between two different kinds of non-rational associative psychological processes: one kind is *brutely* associative – Dill's example is a case where writing down "male" on the pre-test form reminds one that one forgot to put a letter in the *mail* that morning, which then triggers anxiety and distraction which causes one to do badly on the test. The other kind is a sort of *would-be-reason* process, in which one has a thought that *would* "serve as a reason for the elicited behavior" (e.g. concern, double-checking, etc.) "if the alief were converted to a belief" (Dill, p.c.). So one candidate account of what's happening here is that the awareness of one's gender (race, etc.) *would* serve as a reason (in the internal sense, where it makes sense to talk about having *bad* reasons) for concern, lack of confidence in one's responses, double-checking, etc. if one *believed* the negative content of one's stereotype-concordant aliefs. I think that the distinction is well-drawn, and important to notice. I also wouldn't be surprised if the actual facts about stereotype threat turned out to be mixed – that both of these kinds of rogue psychological causation were happening in actual cases of decline in performance under stereotype threat.

(Though maybe that's all Gendler had in mind with the easons talk – this idea of a psychological process that has downstream effects that aren't rationally mediated in the paradigmatic kind of way.)

3. The centerpiece of Gendler's discussion, and probably the most potentially contentious bit of the paper, is the problem that she leads and closes with. Gendler argues that, as members of a society in which objectionable racial attitudes and stereotypes are deeply entrenched, we face a dilemma: either we go on using the objectionable categories, and make ourselves vulnerable to bad epistemic effects from using the problematically loaded categories, as well as expending limited cognitive resources on avoiding and counteracting such effects, or else we refrain from using the problematic categories, which forces us to ignore potentially useful base rate information. As she puts it, "if you live in a society structured by racial categories that you disavow, either you must pay the epistemic cost of failing to encode certain sorts of base-rate or background information about cultural categories, or you must expend epistemic energy regulating the inevitable associations to which that information – encoded in ways to guarantee availability – gives rise." (Gendler, this volume, PAGE REF – p4 of 32 in word doc.) Either way, we're stuck paying an epistemic price for the racism of the society in which we live.

(Maybe this is really better thought of as a trilemma: We can use the categories unreflectively, and wind up with a bunch of bad stereotype-concordant inferences, judgments, attitudes, etc. Alternatively, we can use the categories, but spend a bunch of cognitive resources suppressing or immediately excising the bad stereotype-concordant inferences, judgments, attitudes, etc. Finally, we can avoid using the categories, and fail to code up the base rate information. Also, our situation is pretty clearly not best thought of as a choice between all-or-nothing versions of these three options. It might be psychologically impossible to avoid some of the first type of effects if we use the categories at all, and it might be psychologically impossible to give up the use of the categories altogether. The bad situation, then, is that in any given case, there's no way to avoid paying at least one of these costs. We may, in many cases, wind up paying some mixture of these different costs, and across our epistemic careers, we may pay each, in different mixtures, on different occasions. (Thanks to Anne Barnhill for both of these points.))

I'll spend most of this section looking at some candidate proposals for how to defuse the dilemma. Mostly they're ways of blunting the "don't use the category" horn, and resisting the charge that this will always be accompanied by (or by a risk of) an epistemically costly sort of base rate neglect. I think that they ultimately don't succeed, but it's worth exploring why not.

Gendler talks about the epistemic hazards of "prevalent yet disavowed categorization schema[s]" (this volume, PAGE REF). There are different ways in which a categorization schema could be deserving of disavowal, and on several of ways of understanding what's wrong with the categorization schema, there are some (at least superficially) attractive things to say about how to defuse Gendler's epistemic dilemma.

I'll start by rehearsing a familiar distinction, between *concepts* on the one hand and the *categories* that we pick out with them on the other. (I won't, thankfully, be relying on any very specific way of understanding the distinction, or of characterizing the nature of concepts in what follows – all that's important is the familiar distinction between the category picked out and the bit of linguistic or cognitive representational apparatus that we use to pick it out.) We can then distinguish two ways for a categorization schema to be objectionable – it could be objectionable because there's something wrong with the categories, or it could be objectionable because there's something wrong with some particular concepts that we use to pick out the (themselves unobjectionable) categories. (A third possibility is that *both* the categories and (at least some of) the concepts we use to pick them out are problematic. More on this possibility shortly.)

If what's gone wrong in our social environment is just that we have some racial *concepts* – some ways of singling out racial categories – that are problematic (because, for example, they're loaded up with objectionable stereotypes, etc.), then there's an attractive response available for denying that we're stuck with the sorts of unavoidable epistemic costs that Gendler says we're stuck with.

So long as the underlying *categories* are unproblematic, and there are other concepts available that we could use to pick them out, one defusing strategy is to insist that, since the problem is really with certain racial *concepts*, what we're really called upon to do is simply to refrain from using the particular concepts, or particular terms, which have been loaded up with objectionable racist content and/or associations. It's fine (the defuser might suggest) to use the *category*, but we just have to use it under a different description. If this is our situation, then we can avoid the bad effects, while still coding up the potentially useful base rate information.

This is a proposal that seeks to assimilate the problem to the problem about slurs, on one way of understanding what's happening with slurs. (See for example (Anderson & Lepore, n.d.; Horn, 1989; Hornsby, 2001; Leslie, n.d.; Richard, 2008; Williamson, 2009) for a range of candidate views about slurs and pejoratives.) If what we have is just a relatively tame sort of slur problem, where what we need to do is stop using the objectionable term (concept) for the members of the slurred class, and use a non-slur term (concept) for the class instead, then we can defuse the second horn of Gendler's dilemma – there's no need to go in for base-rate neglect. We just need to encode the base rate information about the category under a different, non-problematic concept or mode of presentation.

There are, I think, three problems with this proposal.

First, it might be that it's the *categories*, and not (or not just) the concepts under which they're picked out, that are problematic. This could be either because it's the categories, rather than the concepts, that have been loaded up with objectionable stereotypes, associations, and aliefs, or because the category is intrinsically objectionable or problematic. (More on this second possibility later.)

Second, we have no guarantee that there will *be* an alternative, uncontaminated concept available. This could be either because there's no alternative available, or because all of the alternatives are themselves contaminated.

The third problem (one which blurs into the second possibility above) is that the problematic concepts could be (and some of the experimental evidence Gendler cites suggests that they are) very available, and very likely to be activated just by thinking about the *category*, under whatever initial mode of presentation.[3] If the problematic concept and associated aliefs are sufficiently available (perhaps in virtue of its social prominence), it's liable to be difficult or impossible to mark that category without calling the contaminated concept to mind in an alief-inducing or alief-activating way. If that's the case, then even if we set out to use some alternative, uncontaminated concept to pick out the category, we would still, at some not-very-cognitively-penetrable level, be operating under the influence of the objectionable concept. (This problem could be made worse by "ironic rebound" effects, where deliberately seeking to avoid the use of the problematic concept calls it to mind ((Huebner, 2009), (Macrae, Bodenhausen, Milne, & Jetten, 1994; Wegner, Schneider, Carter, & White, 1987).)

In practice, it will probably be difficult to pull these possibilities apart. Objectionable associations attached to a particular, highly available concept are liable to bleed over to the category itself, and/or to other concepts that pick out the same category, and they're also liable to manifest themselves when the concept is, because very available, brought to mind by other ways of picking out the category. Gendler's evidence suggests that one or more of these things is in fact pretty systematically happening with the racial categories that are prevalent in our own society.

So it seems as if what we've got isn't just an instance of the tame version of the problem about slurs, where all we have to do is make sure that we pick out the category by means of an unobjectionable term or concept. It's worse than that, because there's no way of picking out the category that won't, or won't be likely to, activate the objectionable aliefs. (Though maybe the slur problem is just as bad, once the slur works its way deeply enough into the cultural background, so that even if it's not actually uttered or explicitly brought to mind, the objectionable aliefs associated with the slur still get activated whenever we think or talk about the category.) It's not just that there's some expression or concept that, in addition to having some unobjectionable descriptive content, also carries the objectionable associations. It's that we're stuck, on account of being in a society with some strongly coded up background stuff, with having the objectionable stuff attached to *any* way of carving reality at that joint.

Another defusing strategy is to maintain that the problematic racial categories are *epistemically* defective, because the categories *don't* carve reality at a joint. According to this proposal, the categories in question are systematically arbitrary, unnatural, gerrymandered, unprojectible, and/or otherwise lousy for reasoning. So just on epistemic grounds, we ought to abandon the use of the contaminated categories. In this case, we might defuse the second horn of Gendler's

[3] As Brendan Dill pointed out in comments on an earlier draft of this paper, the studies on the cognitive costs of interracial interaction seem to speak in favor of this possibility, since the subjects weren't primed with any particular concepts – they were just presented with an interlocutor who was a member of the relevant category.

dilemma by saying that there's no cost to failing to encode the base rate data using racial categories, just as there's no epistemic cost to failing to encode base rate data about grue things or emeroses.

On this view, it's clear what we're called upon to do: give up (to the extent that it's psychologically possible) the use of the category. And this non-categorization won't be cognitively costly, because it's a lousy, unprojectible category. What we should seek to do, on this account, is replace the objectionable categories with alternative, unproblematic categories. And since the objectionable racial categories are highly unnatural, unprojectible, etc., there will always be another category that would do better for doing base rate reasoning with.

There are at least two problems with this response.

First: There's no guarantee that the contaminated categories will always display this kind of unnaturalness, arbitrariness or unprojectibility. It's extremely plausible that in at least very many actual cases, the objectionable categories *are* epistemically objectionable on just these kinds of grounds. But it certainly *could* happen that we take a respectably projectible (or not-completely unprojectible) kind concept, and load it up with a bunch of bad stereotypes and so forth, and get into a situation where members of our society have systematically got a bunch of problematic aliefs about the category. And then, since the category we started out with was a projectible one, we'll be in a situation where giving up using the category, and not coding up the base rate data about it, really *would* be an epistemic cost. Again, the worry isn't that we have good reason to think that our actual racial categories are like this – it seems extremely likely that our actual racial categories really are highly unnatural. But it would be nice to have a response that was sufficiently general to apply even in cases where the categories started off in epistemically good order, and the availability of which didn't hinge on how the empirical facts here play out.

A second, and perhaps more serious, concern about this proposal is that it could happen that presence of the problematic aliefs around the category brings it about that a formerly epistemically useless category *becomes* potentially epistemically useful, once it's been in effect for a while. One thing that could happen is that, because members of the society are sorting people using a highly unnatural, gerrymandered category, and treating them differently based on their membership in that category, the category *becomes* significant, and potentially useful for constructing theories about the social structure of the society. It could quite easily happen that, after an objectionable and initially highly unnatural and unprojectible category has been in use for a while, with people receiving different treatment and different opportunities on the basis of their membership or non-membership in the category, the category comes to be one that could support potentially interesting social-scientific generalizations. (There might, for example, come to be social-scientifically useful generalizations about income, social status, health, etc. that track the boundaries of the category.)

So we could get categories that start off totally gerrymandered, and then become potentially interesting and useful for inquiry *because of the fact that they're loaded up with the bad stereotypes, associations, and so on*. The prevalence of social injustice, etc. associated with the category is liable to give rise to social phenomena

such that, in thinking about those social phenomena, it's potentially useful to mark the line that's drawn by the objectionable concept. (Indeed, it's likely to be indispensible in the process of identifying, pointing out, and combating the negative stereotypes and the resulting social phenomena. Thanks to Brendan Dill for emphasizing this point.)

By way of summary of this section so far, we can illustrate the problem by discussing a certain sort of bad possibility: We get stuck with a social category concept C*, picking out a social category C, early on in our lives. We also get stuck with a lot of lousy, mistaken, and objectionable generalizations about the category, such that we've got a non-short-circuitable process of inference going from application of C* to application of objectionable D*, E* and F*. We then can't just disassociate C* from D* etc. later on, since the associations are happening at a level that's not (or not very) responsive to paradigmatic processes of rational belief revision. One thing we could try is to stop using C*. But this isn't guaranteed to do the trick – maybe C* is so available, that any way of thinking in terms of C is going to activate C*, and trigger the bad associations and inferences. Another thing to try: stop thinking in terms of C. But maybe C is a significant category. Maybe C started off marking a real distinction, and then got loaded up with a bunch of extraneous garbage. Maybe C started off gerrymandered, unnatural, and scientifically uninteresting, but then the social prevalence of categorizing people using C* has brought it about that C is a useful category for thinking about the structure of the society. This is particularly likely if the extant social structures have come to be informed by the fact that lots of people think in terms of C* and therefore people tend to be treated differently, have different experiences, opportunities, etc. in virtue of being members of C than do non-members.

Here is a moral: You can get a society into a position where you can't just get off the hook by using one, rather than another, mechanism for picking out a given category of people. Drawing that line unavoidably (or nearly unavoidably) has got bad cognitive or conative effects, either because we lack any genuinely unproblematic concepts to pick out the category, or because the problematic concepts are so available that they automatically get activated, even when we're trying not to use them. And you're then in a lousy situation. You can try to not draw that line – to not use that category. But there's no guarantee you'll be able to succeed in that task – thinking in terms of that category might also be very deeply ingrained, and not easily short-circuitable. You're also liable to have an ironic rebound, "don't think of a white bear" problem (Huebner, 2009; Macrae et al., 1994; Wegner et al., 1987). And finally, even if you're able to succeed, you've got an unavailability of potentially relevant base rate information problem.

4. Normative conflicts

One thing Gendler is doing in her paper is drawing attention to cases of unavoidable epistemic costs. Another thing she's doing is making a case that we've sometimes got moral reasons to go in for epistemically suboptimal conduct, by foregoing potentially epistemically useful base rate information for the sake of avoiding the use of morally objectionable categorization schemes (or at least that

we ought to take seriously the possibility that this is so).[4] I'll close by drawing attention to some other places where philosophers have argued that we see the same kind of normative conflict. (That is, the sort of conflict between epistemic and some other kind of norms, in which there is an epistemic price to be paid for doing our bit in some other normative domain.)

At a suitably high level of abstraction, the possibility of conflict between epistemic and other norms is old news – at least as old as Pascal.[5] But there are also other, more recent examples, and there's a fair bit of diversity in the grounds for the conflict across cases.

One class of potential cases of this sort of conflict is suggested by Sarah Stroud's and Simon Keller's arguments that friendship sometimes requires forming beliefs, and evaluating evidence, in epistemically sub-optimal ways (Keller, 2004; Stroud, 2006). It seems plausible that something similar is true of our evaluations of our children. It's not just in Lake Wobegon that "all the children are above average" – at least in the minds of their parents. And it seems not at all crazy to think that it would be wrong for things to be otherwise – to think that one is dropping the ball as a parent if one *doesn't* rate one's children more positively than would be supported by a cold-eyed, impartial inspection of the facts.

Even closer to home, there is the case for the pragmatic usefulness of our (very strong, and very difficult, or perhaps impossible, to override) tendencies to overrate ourselves in important respects. As Elga (Elga, 2005) puts it, "There is evidence associating [these] sorts of positive illusions with increased happiness, 'ability to care for others', 'motivation, persistence', and 'the capacity for creative, productive work' (Taylor and Brown, 1988). Furthermore, there is evidence that at least some of the association is causal: that positive illusions help people get by… None of this is surprising. It is reasonable to think that normal (i.e., unrealistically high) self-evaluations promote the sort of self-esteem and self-confidence that help people start projects and persist through difficulties. And it is reasonable to think that a positive self-image makes people happier." (For more on this phenomenon, see for example (Brown, 1986; Brown & Dutton, 1995; Dunning, Meyerowitz, & Holzberg, 1989; Reed, Kemeny, Taylor, Wang, & et al, 1994; Taylor & Brown, 1988; 1994).)

Another possible class of cases is suggested by Julia Driver's discussion of the "virtues of ignorance" (Driver, 1989; 1999). Driver's core case is *modesty*, which she argues requires genuinely underrating oneself, by having a lower opinion of one's own abilities, virtues, etc. than is supported by the evidence.

Perhaps another class of cases (suggested to me by Tamar Gendler – p.c.) is the ones that give rise to the "paradox of choice", (Schwartz 2005) in which having – and knowing about – more options makes one a worse decision-maker. Here, perhaps we have pragmatic reasons not to acquire too much information about our

[4] It's not a sure thing that this is so. It might be, as we saw above, that the categorization schemes are epistemically objectionable as well as morally objectionable, or that their moral objectionableness is due to their bad epistemic effects – of promoting unsound unconscious inferences, for example. But this isn't the only way in which the categorization can be objectionable.

[5] Tamar Gendler (p.c) also points out that this kind of conflict is a common literary trope – for example, one of the things that's so bad about the society in George Orwell's *1984* is the way in which it forces this kind of conflict between, for example, epistemic norms and prudential ones.

available options. (And not just because there are search costs associated with ferreting out the information – rather, because having the information about all of the options available to us, even if we can get it cost-free, is going to hinder our ability to make good decisions, and make us less contented with the decisions that we do wind up making.)

Many of these cases are contentious, and I don't want to try to adjudicate any of the disputes about them here. I just want to draw attention to these cases as some more candidate examples of cases where other kinds of values could trade off against the epistemic ones, and in which there's some temptation to suffer epistemic costs on some other value-based grounds.

This suggests another potential defusing move: maybe we shouldn't be so concerned about paying epistemic costs for our anti-racist commitments. After all, we're already paying epistemic costs for our close personal relationships, for our virtue, and for the maintenance of our self-image and our psychological well-being. This doesn't, of course, challenge Gendler's conclusions – there *is* an epistemic price to pay for living in a racist society. But perhaps it makes the price a bit easier to pay.[6]

References:

Anderson, L., & Lepore, E. (n.d.). Slurring Words. *Noûs*.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals *Journal of Personality and Social Psychology*, *81*(6), 1014–1027. American Psychological Association. doi:10.1037/0022-3514.81.6.1014

Bargh, J., & Chartrand, T. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462–479.

Bortolotti, L. (2010). *Delusions and other irrational beliefs*. Oxford: Oxford University Press.

Brown, Jonathon D. (1986). Evaluations of Self and Others: Self-Enhancement Biases in Social Judgments. *Social Cognition*, *4*(4), 353–376. Guilford. doi:10.1521/soco.1986.4.4.353

Brown, J D, & Dutton, K. A. (1995). Truth and Consequences: The Costs and Benefits of Accurate Self-Knowledge. *Personality and Social Psychology Bulletin*, *21*(12), 1288–1296. doi:10.1177/01461672952112006

Driver, J. (1989). The virtues of ignorance. *The Journal of Philosophy*.

Driver, J. (1999). Modesty and ignorance. *Ethics*.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability *Journal of Personality and Social Psychology*, *57*(6), 1082–1090.

[6] Thanks to Brendan Dill, Tyler Doggett, Jason Stanley, and especially Tamar Gendler, for comments on drafts of these comments, and to Anne Barnhill, Carrie Ichikawa Jenkins, Dilip Ninan, and the participants in 2010 Oberlin Colloquium for extremely helpful discussions.

doi:10.1037/0022-3514.57.6.1082
Egan, A. (2008). Seeing and believing: perception, belief formation and the divided mind. *Philosophical Studies*, *140*(1), 47–63.
Egan, A. (2009). Imagination, delusion, and self-deception. In T. Bayne & J. Fernandez (Eds.). Psychology Press.
Elga, A. (2005). On overrating oneself... and knowing it. *Philosophical Studies*, *123*(1), 115–124.
Fitzsimons, G., & Bargh, J. (2003). Thinking of you: Nonconscious pursuit of interpersonal goals associated with relationship partners. *Journal of Personality and Social Psychology*, *84*(1), 148–164.
Fodor, J. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.
Gendler, T. (2007). Self-Deception as pretense. *Philosophical Perspectives*.
Gendler, T. (2008a). Alief and belief. *Journal of Philosophy*, *105*(10), 634–663.
Gendler, T. (2008b). Alief in Action (and Reaction). *Mind & Language*, *23*(5), 552–585.
Gilbert, D. (1991). How mental systems believe. *American Psychologist*, *46*(2), 107–119.
Gilbert, D., Krull, D., & Malone, P. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601–613.
Gilbert, D., Tafarodi, R., & Malone, P. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, *65*, 221–221.
Horn, L. R. (1989). *A Natural History of Negation*. Chicago: Chicago University Press.
Hornsby, J. (2001). Meaning and uselessness: how to think about derogatory words. (P. French & H. Wettstein, Eds.)*Midwest Studies in Philosophy*, *25*, 128–141.
Huebner, B. (2009). Troubles with Stereotypes for Spinozan minds. *Philosophy of the Social Sciences*.
Keller, S. (2004). Friendship and Belief. *Philosophical Papers*, *33*(3), 329–351. doi:10.1080/05568640409485146
Leslie, S.-J. (n.d.). The Original Sin of Cognition. *Journal of Philosophy*.
Lewis, D. (1982). Logic for equivocators. *Noûs*.
Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound *Journal of Personality and Social Psychology*, *67*(5), 808–817. American Psychological Association. doi:10.1037/0022-3514.67.5.808
Mandelbaum, E. (2011). *The architecture of belief: An essay on the unbearable automaticity of believing*. (J. Prinz, Ed.). University of North Carolina.
Reed, G. M., Kemeny, M. E., Taylor, S. E., Wang, H.-Y. J., & et al. (1994). Realistic acceptance as a predictor of decreased survival time in gay men with AIDS *Health Psychology*, *13*(4), 299–307. doi:10.1037/0278-6133.13.4.299
Richard, M. (2008). *When truth gives out*. Oxford: Oxford University Press.
Schwartz, B. (2005). *The Paradox of Choice: Why More is Less.* New York: Harper Perennial.
Schwitzgebel, E. (2001). In-between Believing. *The Philosophical Quarterly*.
Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*.
Stalnaker, R. (1984). *Inquiry*. Cambridge, MA: MIT Press.

Sterelny, K. (2003). *Thought in a hostile world: The evolution of human cognition*. Oxford: Blackwell.
Stroud, S. (2006). Epistemic Partiality in Friendship*. *Ethics*.
Taylor, S. E., & Brown, Jonathon D. (1988). Illusion and well-being: A social psychological perspective on mental health *Psychological Bulletin*, *103*(2), 193–210. doi:10.1037/0033-2909.103.2.193
Taylor, S. E., & Brown, Jonathon D. (1994). Positive illusions and well-being revisited: Separating fact from fiction *Psychological Bulletin*, *116*(1), 21–27. American Psychological Association. doi:10.1037/0033-2909.116.1.21
Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression *Journal of Personality and Social Psychology*, *53*(1), 5–13. doi:10.1037/0022-3514.53.1.5
Williamson, T. (2009). Reference, Inference, and the Semantics of Pejoratives. In J. Almog & P. Leonardi (Eds.), *The Philosophy of David Kaplan*. Oxford: Oxford University Press.