Reading The Bad News About Our Minds[*]

Penultimate Draft

Nico Silins

Cornell University

Introduction

Psychologists and neuroscientists have delivered a lot of bad news about the inner workings of our minds, raising challenging questions about the extent to which we are rational in important domains of our judgments. Consider for instance the potential influence of evidentially irrelevant factors such as the framing of a question on our moral judgments (Sinnott-Armstrong 2006), or the potential insidious influence of racial biases on our perception and resulting perceptual beliefs (Siegel 2013, 2016, forthcoming).[1] In such cases it might seem that the affected beliefs cannot be rational, or at any rate not as rational as they otherwise would be without the questionable histories they have.

Now, psychologists have received their own bad news as well in the form of the "replication crisis", raising hard questions about the extent to which striking experiments can be replicated and about the merits of the statistical methods used in them.[2] I'll bracket such worries here---you can think of this as an exercise in preparing for the worst. I will instead focus on a central case of an unsettling effect on our perception, and primarily aim to establish that there actually is no impact from it on the rationality of our perceptual beliefs.

[*] I'm grateful for the help of Jessica Brown, Carolina Flores, Jack Lyons, Matt McGrath, Susanna Siegel, Lu Teng, and Jonna Vance, as well as audiences at Bled and St Andrews.

[1] For some further relevant skeptical discussions, see Carruthers (2011) or Schwitzgebel (2008, 2011) on introspection, or Doris (2015) on reflective agency.

[2] For sample overviews, see the introduction of Francis 2016 or section 1 of Machery 2019.

To reach my goal, I will start with a rough review of different ways bad news about our minds might negatively affect our rationality. I'll then propose a test for when negative consequences would follow from the truth of some bad news about our minds, and I will argue that the test is not passed in a paradigmatic case of influence of bias on perception. I will then close by looking at potential wider implications of the psychological bad news for debates about internalism vs. externalism in epistemology.

1. Framework

I'll begin with a survey of some standardly recognized ways in which some evidence or a fact can negatively impact the rational standing of our beliefs. To keep things simple, I will mainly focus on our beliefs rather than our more fine-grained levels of confidence, and I won't try to provide any sort of formal model (for a sample attempt, see Kotzen 2019).

First consider **ordinary defeaters** of a source of your belief. Suppose that, at noon, you believe that it is there is a high chance of rain this afternoon, and believe this on the grounds that your friend, who you know to be generally trust-worthy, said so. One standard example of an ordinary defeater here would arise if you checked your phone at 1pm, and gained the evidence that the latest weather update says that there is a low chance of rain this afternoon. A standard example of a different kind of ordinary defeater here would arise if you gain the evidence at 1pm that your friend is happy to lie if necessary to get out of going to a picnic. In both cases, you do gain evidence at 1pm that reduces the level of rational support you have from your friend's testimony at 1pm. However, you arguably do not gain evidence at 1pm that your belief wasn't rationally supported all the way back at noon. At any rate, it is not built into being an ordinary defeater that such an impact would take place.

In general, an ordinary defeater is evidence of yours that reduces the level of rational support you gain from a source, without necessarily conflicting with your

having gained rational support from the source prior to acquiring the defeating evidence.

Now forget about the evidence a person has, and instead consider the potential impacts of facts that might not be included in their evidence. For a common example, suppose that Harold has tried to reason carefully to solve a complicated logic puzzle, but in fact has been under the influence of reason-distorting drugs while doing so, and has in fact reasoned fallaciously. Harold has no evidence of being in such a condition, so we do not have any ordinary defeaters here to speak of, but nevertheless his belief in the solution of the logic puzzle plausibly fails to be rational. The fact that he has reasoned fallaciously is plausibly a **destroyer**, a fact that makes it the case that his belief in the solution of the puzzle is not formed on the basis of a source that gives it any rational support (destroyers are also known as "**blocking debunkers**" in the terminology of White 2010: 475).

You might think that the reason-distorting drugs do not make Harold's belief entirely bereft of support. In such a scenario we would merely have a **damager**, very roughly a fact that makes his belief have less rational support from his reasoning than his belief otherwise would, while still leaving room for his belief to have some rational support from his reasoning.

2. A test for being a destroyer

Now that we have a grasp on defeaters, destroyers, and damagers, I will propose a test for being a destroyer. I will keep damagers in the background for now, but we will return to them later.

My core idea is a special case of a general rule: once you have hit rock bottom, you can't go down any further. In particular, you can't defeat what's already been destroyed. For example, if some bad news from psychology rules out that someone has rational support from moral intuition for a moral belief, then learning the bad news cannot worsen the rational standing of their moral belief derived from moral intuition. Since any potential level of support from moral intuition will be at rock bottom, it is not possible for the belief now to acquire even less support from

moral intuition. (For all we have said, the overall rational standing of the belief might able to decline further, but our focus is specifically on how there is no room for any further decline of support from moral intuition.) More generally:

> (Test): If the fact that F is a destroyer of rational support from a source S for someone's belief that P, learning that F doesn't worsen the rational standing of their belief that P derived from source S.

To understand Test and its appeal, it might help to consider the following example. Daewon initially formed a belief that it is likely to rain this afternoon, by his lights on the basis of Emily's testimony that it is likely to rain this afternoon. But Daewon subsequently learns the following fact: Emily never said that it is likely to rain this afternoon, and Daewon didn't even have an auditory misperception of her saying it is likely to rain this afternoon—he instead formed the belief she said that on the basis of his own wishful thinking. The fact about the origin of his belief is a destroyer. And the rational standing of Daewon's belief derived from Emily's testimony doesn't go down as a result of his learning, instead he simply learns that his belief had no support from her testimony to begin with.

When I apply Test to various cases, my crucial claim will in effect be that the given piece of bad news from psychology does act as a defeater rather than as a destroyer. More generally, I will reason along the following lines:

Template
(Template for Fail): Learning that F worsens the rational standing of your belief that P derived from source S.
(Template for Test): If the fact that F is a destroyer of rational support from a source S for your belief that P, learning that F doesn't worsen the rational standing of your belief that P derived from source S.
So,
The fact that F is not a destroyer of rational support from a source S for your belief that P

In formulating the test in terms of "learning", I have in mind a condition that requires it to be the case that F, such as coming to know that F. Formulating the test in this way ensures that a given destroyer obtains when the test is applied to that

destroyer. Now, you might in principle gain misleading evidence in favor of its being the case that F when in fact it is not the case that F, or simply gain evidence in favor of its being the case that F in an inconclusive way compatible with its not being the case that F. Here I will leave open what effect such evidence has, if any, on the rational standing of your relevant belief in such scenarios.[3]

We need to consider an important complication. Given the room for debate about various claims by contemporary psychology, we most likely at best have evidence in favor of this or that piece of bad news from psychology, without yet have learned or come to know that the bad news is true. When I apply my test in various scenarios, I will assume that we have come to learn the candidate for being a destroyer obtains. My crucial move will be to hold that in those scenarios, the rational standing of our belief from the relevant source does worsen. The candidate for being a destroyer will thereby fail to be a destroyer in the case. Proceeding in this way will also allow us to conclude that a given candidate is not actually a destroyer even given our current inconclusive evidence, since actually being a destroyer would entail that, if one did learn that the candidate obtains, one would not thereby degrade the rational standing of the relevant belief.

3. Application

I'll now apply our test to the current debate about the epistemic import of some forms of so-called cognitive penetration of perception.

I'll set the stage by reviewing background about what it would even be for our perception to be subject to cognitive penetration (drawing extensively on Macpherson 2012 and my 2016).[4]

The key broad idea for present purposes is that, if our perception is subject to cognitive penetration, then we sometimes perceive the world as being a certain

[3] For further discussion of such cases, see Christensen 2010, Lasonen-Aarnio 2014, Schoenfield 2018 or Weatherson 2019

[4] For some alternative approaches to setting up the issues, see Stokes 2013, Shea 2014, or Gross 2017.

way as a result of how we antecedently expect the world to be, where those expectations themselves may not be rational to have. For the sake of simplicity, I will usually focus on the role of our expectations, but there is plenty of room for the possibility of "cognitive" penetration by desires or other states (see Stokes 2012). The crucial point is that the relevant states influencing our perception fail to be states built into our perceptual system. Critics such as Firestone and Scholl (2016) deny that our perception is ever the result of cognitive penetration, but can still coherently allow that our perceptual system works partly by using built-in assumptions about the world, for example an assumption that illumination generally comes from above.

Not just any effect of our expectations on our perception is enough to generate a case of cognitive penetration. Some influences of our expectations on our perception might be filtered through our attention in a particular way that presumably shouldn't count, at least if we are aiming to isolate a potential perceptual phenomenon of theoretical interest (Fodor 1984, Macpherson 2012). For example, you might expect someone to come around the corner and thereby direct your attention to that part of the hall, and in turn thereby perceive that part of the hall as a result of your expectation. This won't be enough to count as cognitive penetration. Notice that even if another perceiver failed to expect anyone to come around the corner, other things being equal they would still perceive the world the same way as you, were they to attend to the same location as you.

In order to get a proper example of cognitive penetration in mind, it is better to compare a pair of perceivers, holding fixed their spatial attention, while varying their perception. For example, consider the effect of so-called "memory color", and how someone with antecedent expectations about the blueness of Smurfs might see a grey-scale image of a Smurf as slightly blue (as in Witzel et al 2011). Extrapolating from the study, we presumably could compare a perceiver entirely unfamiliar with Smurfs, engaging their spatial attention on the same kind of grey-scale figure in the

same way, where the uninformed perceiver would instead just see the grey-scale image as being grey-scale.[5]

Generalizing from the current example, we might think that perception is cognitively penetrable just in case the following scenario is possible:

> two people are the same with respect to their sensory inputs, the state of their sensory organ, and the orientation of their spatial attention, and they are still different with respect to what their perception is like, because of their beliefs, desires, or other cognitive states.[6]

The problem is that while we have ruled out some irrelevant effects of our expectations on our spatial attention, there is more work left to do to get a plausible sufficient condition for cognitive penetrability. Two people might attend to the same location, but still attend to different properties due to differences in their cognitive states, for example differing interests in hue vs. saturation. If such antecedent non-perceptual differences result in differences in what their perception is like, it is not yet clear whether we have enough for a case of cognitive penetration. In particular, one possibility is that attention merely places a filtering role with respect to what properties the pair perceive, with A perceiving the Fness of an object thanks to attending to the Fness of the object, and B failing to perceive the Fness of the object due to inattention to the Fness of the object. It is not clear whether such a merely selective role of attention should suffice for an interesting case of cognitive penetration.[7]

A tempting solution to the problem is to hold fixed visual attention entirely across our compared perceivers, not just spatial attention. The new proposal would be that perception would be cognitively penetrable just in case the following scenario is possible:

---

[5] For a range of related studies, see Hansen et al., 2006; Kimura et al., 2013; Olkkonen, Hansen, & Gegenfurtner, 2008; Witzel, Valkova, Hansen, & Gegenfurtner, 2011.

[6] I adapt this formulation from Macpherson 2012, but use the notion of perception rather than of experience.

[7] For an example of this point in the literature, see Firestone and Scholl 2016. For critical discussion of the point and of how strong a conclusion can be drawn from it, see Lupyan 2015, Gross 2017 or Green forthcoming.

> two people are the same with respect to their sensory inputs, the state of their sensory organ, and the orientation of their visual attention, and they are still different with respect to what their perception is like, because of their beliefs, desires, or other cognitive states.

While our previous attempt asked for too little, the current attempt asks for too much. While not just any effect of non-spatial attention on perception should be ruled in as a case of cognitive penetration, we arguably should leave room for some effects of attention on perception to count as cases of cognitive penetration. For example, consider Green forthcoming's discussion of experiments by Ling et al 2009. These experiments seem to show that, when we voluntarily guide our attention to a cued direction of motion, that dramatically improves our ability to perceive the global direction of motion of a group of dots relative to the cued direction of motion. Here our voluntarily guided attention seems to increase the perceptual discriminability of a feature, not playing a mere role of selection. In order to understand the differences in perception that result from such potential cases of cognitive penetration, where our volition shapes our perception via attention, we need to compare perceivers who are different with respect to attention rather than the same.

We need more work to pin down a proper formulation of what it would take for there to be theoretically significant cognitive penetration of perception. In what follows I will take it that we can get by with our rough and ready grip on the phenomenon, in particular using the example of memory color from Witzel as a paradigm.

Assuming now that cognitive penetration does sometimes happen, there's plenty of room for debate about what sort of epistemic repercussions it might have. As figures such as Churchland have suggested from the beginning, the upshot of at least some cases of cognitive penetration could be a matter of enlightening expertise rather than any form of epistemic threat. For an extreme example from Churchland, consider the bulging eyes of our descendants:

> They do not feel common objects grow cooler with the onset of darkness, nor observe the dew forming on every surface. They feel the molecular KE of common aggregates dwindle with the now uncompensated radiation of their energy starwards, and they observe the accretion of reassociated atmospheric H20 molecules as their KE is lost to the now more quiescent aggregates with which they collide . . . (1979: 29-30)

Churchland presumably would not hesitate to say that our descendants get knowledge or rational belief from such cases of perception, however suspicious the rest of us might be about the possibility of the striking scenario he describes (see e.g. Fodor 1984).[8]

To consider a much less extreme example, it is now something of a commonplace that an expert radiologist might see more than the rest of us when looking at an x-ray, and thereby gain rational support for their beliefs from their perception in a way not yet available to the uninformed. Whether this is really a genuine case of cognitive penetration is however not entirely clear. Rival hypotheses still need to be ruled out: perhaps the radiologist simply has a better sense of where to selectively attend in the x-ray, or perhaps can better form judgments on the basis of seeing exactly what the rest of us see.

However, for another source of potential examples of beneficial cognitive penetration, consider the experiments of Ling et al we just considered where our attention seemed to increase the perceptual discriminability of a feature of the world. Assuming that these are cases of cognitive penetration, they also seem be cases of enhancement of our ability to gain knowledge or justified beliefs from perception. Even if the participants were shaping perception through voluntary direction of their attention, the ability of their perception to teach them about direction of motion seems if anything to be improved.

In saying that the potential cognitive penetration of perception by attention so far looks epistemically good rather than bad, I depart from the recent assessment by Wu 2017, who opens his paper with the following:

[8] It is especially odd that, while the eyes of Churchland's children bulge with scientific detail, those eyes are blind with respect to ordinary categories such as dew.

> I argue for cognitive penetration where the epistemic consequences are actual: one's beliefs depend on what one notices or attends to, but attention is biased by cognition. Such cognitive influence raises pervasive challenges for any epistemic agent trying to discover the truth (2017: 5-6).

Given the actual cases of attention and cognitive penetration Wu surveys, the worry he expresses has yet to have bite. One family of Wu's perceptual examples involves a series of experiments by Desimone et al. The experiments involve a task for macaques of matching a delayed stimulus to a previous target, in which their neural response to a stimulus is reduced by a lack of attention to the stimulus. The authors tentatively conclude that "attention gates visual processing by filtering out irrelevant information (1985: 784)." Setting aside the details of the experiments, this sounds like a boon rather than a challenge to an agent investigating the world. Wu's other central example comes from Yarbus' work tracing out how eye movements are shaped by one's intentions to remember aspects of a scene in a painting. This example also does not seem to involve any epistemic impairment of perception, since attention here does not distort perception in any way, nor is attention directed by cognitive states that are irrational in any way. I don't think Wu has yet given any evidence of an actual epistemic hurdle for perception raised by attention.

Let's now put issues about attention into the background, and turn to more insidious forms of cognitive penetration that are much harder to see as forms of enlightenment. The recipe here is to consider antecedent cognitive states that are themselves rationally suspect, such as cases of wishful thinking, or racial prejudices, or implicit biases (Markie 2005, Siegel 2012, 2018, forthcoming, McGrath 2013, Vance 2014). The view is that, when we perceive the world as being a certain way as the result of cognitive penetration by such rationally suspect states, and form a belief that the world is that way by taking our experience of the world at face value, our resulting belief that the world is that way is not rational. In what follows I will address this claim as "the negative view". (Notice that the negative view need not be driven by the claim that our perception has ended up misrepresenting the world.

For all we have said, someone might luckily have accurate perception that results from cognitive penetration, but still fail to get a rational belief from perception because of the way their irrational non-perceptual states were involved in that cognitive penetration).

I will now present my argument against the negative view, starting by setting out a central case to consider. There is sadly an abundance of potential examples of problematic cognitive penetration to choose from. To pick just one area, psychologists such as Keith Payne, Jennifer Eberhardt and Joshua Correll have done many studies of how perceivers primed with images of Blacks can misidentify phones or tools as guns (see Wittenbrink 2019 for an overview). I will focus here on a different kind of study by Hugenberg and Bodenhausen 2003 of emotion perception, one that has been less widely discussed by philosophers.[9]

Their main experiment had White American subjects view animated "morphing videos" showing faces transitioning from angry expressions to happy expressions, as shown in fig. 1.

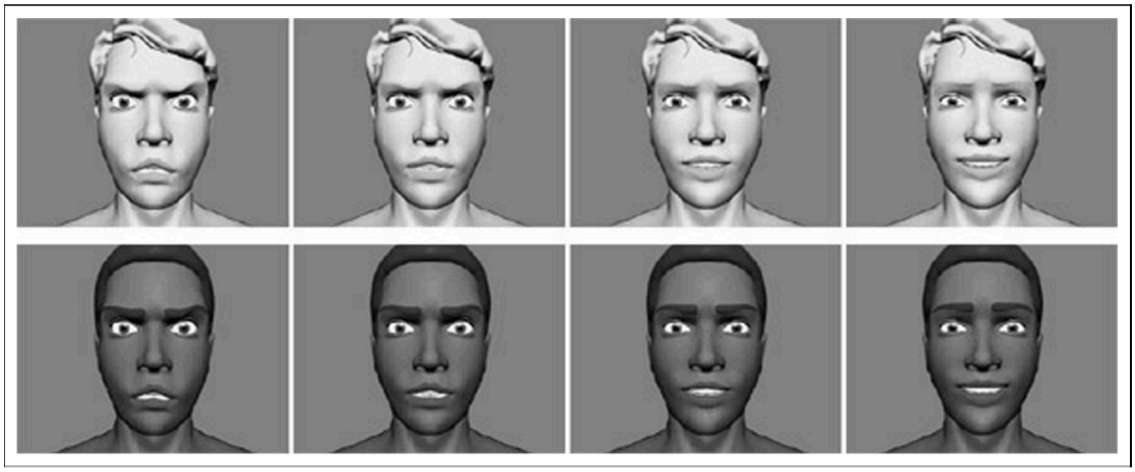

Fig. 1. Four frames of one angry-to-happy movie with the White (top) and Black (bottom) target faces. The figure shows gray-scale reproductions of the original color images.

[9] For some studies with closely related results, see Hugenberg 2005, Hutchings et al 2008, or Bijlstra 2010.

Hugenberg and Bodenhausen's key finding was when the viewing White subjects scored high on an implicit association test linking Blacks with unpleasantness, those White viewers maintained their judgment that the Black face was angry for roughly a second longer than with White faces.

In my use of the study in my argument, I will use the following assumptions about it. First of all, I will assume that the effect here is genuinely on how the White subjects perceive the Black faces, rather than merely on how White subjects update their judgments or beliefs on the basis of their unaltered perception. Second, I will assume that the effect here comes from an irrational bias or prejudice on the part of the White subject. I leave open the exact character of the influencing state, the crucial points being its irrationality and its failure to be part of the perceptual system.[10] These assumptions only favor my opponent, since they allow us to work with a case of an effect on perception that is clearly problematic in some way.

I now ask us to consider a White subject John who learns that his perception of anger is a result of cognitive penetration by his racial bias or prejudice [more briefly, who learns the fact that P]. My crucial claim here is the following:

(Fail): Learning that P worsens the rational standing of John's belief that the face is angry derived from John's perception.

My view is that this is the plausible judgment to make about the case. Contrast a judgment that, when John learns that his perception is a result of cognitive penetration by his racial bias, there is no impact on the rational standing of his belief derived from perception. This is the judgment that is predicted to be true by the negative view, but I take it to not be a plausible judgment to make about the case. (The negative view does not predict that when John learns the bad news, he learns that his perceptual belief was not rational to believe with---we have left open what opinion John himself has about the effects of insidious cognitive penetration).

[10] For recent discussion of the nature of implicit bias, see Mandelbaum 2016 or Brownstein 2018.

If you find yourself on the fence about Fail, it might help to compare John's scenario to a pair of cases of testimony. In both of them, Daewon has initially formed a belief that it is likely to rain this afternoon, by his lights on the basis of Emily's testimony that it is likely to rain this afternoon. In the first variant, that we have already seen as an example of a destroyer, Daewon goes on to learns that Emily never said that it is likely to rain this afternoon, and he didn't even mishear her as saying it is likely to rain this afternoon—he instead formed the belief she said that on the basis of his own wishful thinking. In the second variant, Daewon goes on to learn that Emily's testimony that it is likely to rain this afternoon was driven by her hidden agenda against having a picnic. Here I would say we have a classic example of defeat, where the previously high rational standing of Daewon's belief derived from Emily's testimony goes down. I submit that, when we compare what happens when John learns the bad news about his perception, John's position is analogous to Daewon's second case of defeat rather than Daewon's first case of learning about a destroyer.

We can now apply our test for being a destroyer as follows:

(Test): If the fact that P is a destroyer of rational support from John's perception to believe that the face is angry, learning that P doesn't worsen the rational standing of John's belief that the face is angry derived from John's perception.

We may now conclude against the negative view that

(Conclusion): The fact that P is not a destroyer of rational support from John's perception to believe that the face is angry.

While I won't trace out the applications here, or how they might perform more or less well in different domains, our template can also be applied to many other cases of learning bad news from psychology about the workings of our minds.

I'll now turn to a series of objections to my argument.

Objections and Replies

First, one might say that I am directly begging the question against my opponent. While Fail is incompatible with the negative view, Fail is a specific claim about what happens in an individual case of learning about how one's perception was generated, not by any means the direct denial of the negative view. So I am not in any direct way begging the question against the negative view. I am instead arguing against the view by teasing out a consequence of the view that seems to be false.

As a further objection, the opponent might deny that there even seems to be a decrease in the level of support for John's belief from perception. Perhaps the intuitively plausible judgment about the case is instead actually that there is no such decrease (more below about what sorts of other impacts there could be on the rational standing of various beliefs of John's).

Here we would have a Clash of The Intuitions. I do not have a further positive argument in favor of Fail (a kraken) to release. But I should clarify that my goal is not to change the mind of a proponent of the negative view. My goal is to present evidence that is compelling to a neutral onlooker to the debate. I should also clarify that I do not mean to apply a principle to the effect that, if your belief is not rational, then you must be able to figure out on your own that your belief is not rational. Some cases of persistent wishful thinking or of mental delusions seem to show that such principles are false (assuming mental delusions are beliefs).

Another broad line of objection agrees that there is an epistemic disruption when John learns the bad news, but denies that Fail locates the problem in the right place.

On one variant of the objection, the epistemic disruption is not to the rationality of any of John's beliefs. Epistemic appraisal can come in many forms, so the impact of learning the bad news could be on some epistemic dimension other than rationality. For example, perhaps John was initially irrational but blameless in believing that the face is angry, and then ceases to be blameless in having that belief upon learning the bad news.

While I agree that there are multiple forms of epistemic assessment, I think this variant assumes too much confusion in our judgments about John. Saying that we weren't thinking about rationality in our assessment of his case is too much of a change of subject.

On the more promising variant of the objection, learning the bad news does impact the rationality of John's beliefs, just not his belief that the face is angry. In particular, one might claim that I am simply committing a level confusion in my verdict about John's situation (Alston 1986 is a classic source for this sort of point). Perhaps the rational impact of learning the bad news starts only at the second floor, on the rational standing of John's belief that he is rational in believing that the face is angry.

I have two main responses here. First, in order to stick with this objection, my opponent would have to allow for a certain kind of level mismatch. The opponent would need to hold that, before learning the bad news, John was not rational in believing that the face is angry, but still was rational in believing that he was rational in believing that the face is angry. Now, it is controversial whether such mismatches can arise (see Weatherson 2019 or Smithies 2019 for an overview of much of the recent debate about level principles). In particular, it is not clear how the opponent can explain the rationality of John's second-order belief without that explanation trickling down into a prediction of the rationality of his first-order belief. But such a prediction would be in tension with the opponent's main claim about John.

Second, it was not seamless for me to shift the spotlight to the rational status of John's second-order belief, and it may not have been seamless for you. This suggests that we were not initially thinking about the rational status of his second-order belief when thinking about the case, and are not guilty of the confusion identified by the opponent. In any event, once we have brought the level distinction clearly into play, explicitly blocking the potential confusion, Fail still seems to me to be true. So the appeal to level distinctions does not seem to help the opponent.

I'll now turn to a quite different line of objection, one that is especially important since it takes us deeper into the nuance available to proponents of the negative view. The basic move is to try to concede that Fail is true, but understand the negative view about cognitive penetration as compatible with Fail. We can state the move in terms of our earlier distinction between destroyers and damagers, where damagers reduce your level of support for a belief from a source while still leaving some level of support intact. The moderate negative view can be understood as holding only that bad forms of cognitive penetration act as a damager, leaving open whether or not they act as a destroyer. Indeed, Siegel's signature way of stating her view, in this quotation applied to a particular example, is qualified as follows:

> Her experience is merely *downgraded*, meaning its epistemic power is reduced below the baseline (2017: 23).

Setting aside how exactly to unpack "epistemic power" or the relevant baseline, the key point here is that Siegel's thesis is hedged in terms of degrees (McGrath 2013 does not seem to qualify his view in terms of degrees).[11] So perhaps

---

[11] There are some important complications for Siegel's formulation of a negative view in terms of a decrease below a "baseline" level of support from experience.

A natural reading of "baseline" would be as "how much experience normally justifies you for beliefs about the external world". Downgraded cases of experience would then be cases that justify less than whatever that amount is (assuming there is such a thing).

This approach is tricky for cases where experiences are already performing below par for other reasons. When you see someone in the distance, your experience is already downgraded, whether it's a bad case of a cognitive penetration or not. So here it's trivial to say that, if the experience is a bad case of cognitive penetration, then it's downgraded. Also, we need to handle cases of experience that are performing well above par, say of the color of something right in your face, but where those experiences are still negatively affected by cognitive penetration. We need to allow for a negative impact that still leaves something above the ordinary baseline level of justification ordinary experiences provide.

In response, you might relativize the "baseline" to much more specific types of experiences. However, this would be tricky if some experiences have certain contents only through bad forms of cognitive penetration, say experiences with racist contents. All cases of such types of experiences would have an epistemic defect, and you wouldn't understand the defect as a matter of falling below a bar set by the good cases of them, since there are no good cases of them.

Siegel (and others) can simply take on board my claim that the impact of racial bias on John's perception fails to act as a destroyer. On this moderate negative view, bad cognitive penetration reduces the amount of rational support given by perception, but learning about bad cognitive penetration can reduce the amount of rational support given by the relevant perception even further.

My main response is that the objection is unstable: it's not at all clear how to motivate the view that bad cognitive penetration is a damager without becoming committed to the view that bad cognitive penetration is a destroyer.

Consider for instance the way Siegel initially motivates her view. She considers an example in which Jill has a fearful unjustified suspicion that Jack is angry at her, where that suspicion plays the role of generating Jill's perception of Jack as being angry at her through a process of cognitive penetration. In response to her perception, Jill re-affirms her belief that Jack is angry at her. Now, according to Siegel,

> She seems to have moved illicitly from her starting suspicion to a strengthening of it, via her experience (2017: 6).

I take Siegel here to rely on an analogy with viciously circular reasoning (see also the "gossip circle" example of her 2013). Just as you cannot gain rational support for your belief that p by inferring that p from your belief that p, Jill cannot gain rational support for her attitude about Jack on the basis of her experience or perception as of Jack's being angry, when that experience or perception is itself the result of that attitude.

The key problem here for our purposes is that viciously circular reasoning is a source of zero rational support, not a source of somehow attenuated rational support. So if bad cognitive penetration is epistemically deficient because of its analogy with viciously circular reasoning, we should expect bad cognitive penetration to be a destroyer rather than a damager.

An advantage of the view that cognitive penetration is a destroyer is its simplicity.

So far the initial motivation for a moderative negative view takes us all the way to the full-strength negative view that bad cognitive penetration is a destroyer.

Now, Siegel and McGrath 2013 also have a further, much more central line of argument that draws instead on an analogy between bad cognitive penetration and the basing of beliefs on irrational beliefs (a related argument is also defended by Vance 2014, focusing on an analogy with emotion rather than belief). The core idea is that if we look closely at the process whereby an irrational fear, racial bias, piece of wishful thinking or the like generates a case of perception, that process will turn out to share important epistemic properties with inference. Inferring a belief in a conclusion from an irrational belief in a premise, however deductively valid or inductively appropriate the inference might be, does not result in a rational belief in the conclusion. When garbage goes in, garbage comes out---inference does not up-cycle. Now, when you base a belief about the world on your perception, where that perception is itself generated in part by an irrational state of yours, we end up with a chain from an irrational state to your perception to your perceptual belief. Arguably this chain is a process of inference, or at any rate similar enough to a process of inference, so that the chain inherits the epistemic profile of inference. In particular, when an input is an irrational state, the output will not be a rational belief.

For our present purposes, we need not worry about whether the crucial points of analogy with inference hold up. What matters in the present context is that when you infer a conclusion from an irrational belief in a premise, you do not thereby gain any rational support for your belief in the conclusion. Inference from irrational beliefs is a destroyer rather than a damager. So if the present argument from analogy is successful, we should expect bad cognitive penetration to be a destroyer rather than a damager. Again we fail to have a line of argument that would reach a stable stopping point at the moderate negative view.

There is a related reason to suspect that the basing argument does not provide a useful tool to explain why cognitive penetration might impair the ability of perception to rationally support beliefs. Suppose some cognitive penetration indeed does impair the rational support given by perception. If so, this effect must

be able to happen to greater and lesser degrees, for example according to the varying degrees of irrationality of the cognitive states involved in shaping our perception. But the basing argument cannot accommodate this point, since inference from an irrational state fails to result in a rational belief full stop, with no variation of degree in the matter. Your belief in the conclusion is not more irrational just because your belief in the premise was less rational. Now, there might be variation in the degree of quality of various inferences, but that is not the focal point of the basing argument. The basing argument was oriented instead around the around the point that you get zero rational support from an inference from an irrational starting point. So the basing argument is too blunt an instrument to be a good guide to the epistemology of cognitive penetration.[12]

There is one last line of objection I will consider. This one also tries to make the negative view compatible with Fail, except this time relying on a point about the way in which perception gives rational support for belief, rather on any point about degrees of rational support.[13]

When your perception makes it rational for you to believe that p, your perception could do so in a way that is non-inferential, not in virtue of its being rational for you to have this or that background belief, or your perception could do so in a way that is inferential, that does happen in virtue of its being rational for you to have this or that background belief. Now, one might qualify the negative view as the claim that bad cognitive penetration is a destroyer of any non-inferential way for your perception to give rational support for beliefs about the external world, leaving open what happens with inferential ways of getting rational support from perception. Here the negative view is restricted to what you might think of as foundationalist ways of justifying beliefs. So when John initially believes that the face is angry on the basis of his apparent perception of anger, perhaps his

[12] Thanks here to discussion with Jack Lyons. He also pointed out to me that reliabilist approaches such as his own might have a better time capturing degrees of badness in cases of cognitive penetration. I set aside the appraisal of reliabilist approaches here, but argue against them in my 2014.

[13] Thanks to Jonna Vance and Lu Teng for discussions of this line of objection.

perception still does make it rational for him to have that belief, just thanks to his background belief that [if it visually seems to him that the face is angry, then the face is angry]. And then when he learns that his perception is the result of cognitive penetration by a racial bias, perhaps the rational standing of this background belief is lowered, bringing the rational standing of his belief that the face is angry down with it as well. Here we seem to have a coherent story about how learning the bad news about his perception could lower the rational standing of his belief that the face is angry, even if his perception never gave non-inferential rational support for his belief that the face is angry.

I would say the main challenge here is to spell out the relevant background beliefs so that they are both affected by bad news about cognitive penetration, and are had or available to every subject we could use to set up our argument.

The initial version of the objection relied on John's having a background belief that his perception is accurate. But as I mentioned earlier, cases of bad cognitive penetration need not involve any inaccuracy in perception, or even any apparent inaccuracy in perception. The psychologist could ask John whether he wants the good news first, and then inform him of the following conjunction: [the face does happen to be angry, but the way you see it is driven by your bias linking Black faces with anger]. Here there might not be any damage to the overall rational standing of John's belief that the face is angry, indeed there could even be a net gain in the rational standing of his belief. But I think there would still be damage to the way that his perception improves the rational standing of his belief that the face is angry. Instead of gaining support from his perception to believe that the face is angry, he would now gain support from the testimony of the scientist to believe that the face is angry. The objector here wouldn't have an explanation of why Fail remains true of the case.

An alternative is to rely on John's having a background belief that demands more than accuracy. Perhaps John relies on a belief that, if it visually seems to him that the face is angry, he sees the anger of the face. Seeing the anger of the face is a form of perceptual contact that requires more than having a perception that happens to match how the face is, and perhaps is incompatible with John's

perception resulting from bad cognitive penetration. John's perception might instead be something like a hallucination of an angry face that just happens by chance to be accurate. However, given the potential massive extent to which our perception is informed by various top-down effects (see e.g. Clark 2015), it is unclear whether bad cognitive penetration is incompatible with seeing the anger of the face. There is plenty of room for the possibility that, when we successfully make visual contact with the world by seeing it as it is, our perception is nevertheless the result of cognitive penetration. So it's not clear whether the proposed background belief is in any tension with John's reception of information about cognitive penetration resulting in his perception.

A further alternative is that John relies on a belief that, if it visually seems to him that the face is angry, his perception is not the result of bad cognitive penetration by a racial bias. Here we have an excellent candidate for being affected by learning bad news about our minds, but a bad candidate for being a belief upon which we normally rely. We now are considering a highly rarified content that is not believed or available to many subjects who nevertheless remain perfectly rational in relying on their perception. (While it is true that the belief is available to John, I take it to be ad hoc to hold that he relies on this background belief that is not even available to so many others).

There is a classic challenge here for anyone seeking to give an inferential account of how perception gives rational support for belief (see e.g. Burge 2003). Given that we can form beliefs on the basis of perception in a way that is rational, apparently without doing so also on the basis of background beliefs about various properties of our perception, it is hard to cover enough cases with the view that perception gives rational support for belief in an inferential way. Since even unreflective subjects are vulnerable to defeat of their perceptual beliefs upon learning bad news about how their minds work, proponents of the negative view won't be able to explain how such defeat happens in enough cases.

I acknowledge that someone might have rational support for a proposition about their perception, and yet not believe that proposition at all, or not have a perceptual belief partly on the basis of believing that proposition (in other jargon,

they would have "propositional justification" for such propositions). So even when unreflective subjects get bad news about their perception, the rational standing of those propositions could go down. This is all true, but does not help the opponent. We are examining the rational standing of John's actually held belief actually derived from perception ("doxastic justification", in other jargon). This status need not go down even if the rational standing of propositions in the neighborhood about John's perception goes down. Compare: if A has received the same testimony from B and C, but only believes on the basis of the testimony of B, A need not lower her level of confidence when learning about dubious motivations behind C's testimony. As far as John's actually held perceptual belief is concerned, the background propositions highlighted by the opponent are an idle wheel.

Restricting the negative view to a claim only about non-inferential rational support doesn't help.

Summing up, I've argued against the view that cognitive penetration of perception by biases or other problematic states destroys the ability of perception to rationally support belief. I've also responded to several potential attempts to undermine the key premises of my argument, or to qualify the negative view so as to escape the argument.

4. Destroyers and the Internalist/Externalism Debate

But what if I'm wrong? What further repercussions for views in epistemology would there be if proponents of the negative view were right? In this last section, I'll step back to examine the implications of bad news from psychology for classic debates about internalism vs externalism in epistemology, expanding on previous work in my 2018. I'll start by articulating an attractive picture of the implications, and then proceed to explain why that picture is wrong. I'll keep our focus still on the case of cognitive penetration.

If John's perception did fail to give any rational support for his belief that the face is angry, it is tempting to think that would favor an externalist approach in epistemology. After all, on a standard understanding of the internalist/externalist debate, internalists can't assign any role in epistemology to features of our minds that are introspectively invisible to us, and externalists easily can assign an undermining role to features of our minds that are introspectively invisible to us. Against this line of thought, I will argue it relies on a widespread misunderstanding of what the internalism/externalism debate is about—the core of the dispute is actually not about the potential epistemic role of what is inaccessible to us.

I'll first flesh out the tempting line of thought with the earlier example from Siegel of Jill and Jack. In particular, compare Jill with a counterpart Jill+ whose perception is not colored by a fearful suspicion that Jack is angry at her, or otherwise affected by any bad form of cognitive penetration, and who encounters a Jack who is angry at her she perceptually registers as such. Here the negative view would predict that Jill and Jill+ differ in what they have rational support to believe. Nevertheless, Jill and Jill+ are plausibly the same with respect to what it is rational for them to believe on the basis of introspection---the relevant differences in how their perception is generated flies well below their introspective radars.[14] So you might think that the negative view is incompatible with internalism in epistemology, given the privilege given by internalism to what it is rational for us to believe on the basis of introspection.

We can now state the line of thought more generally as follows (see Puddifoot 2016 for an example of it put to work in the literature):

(Internalism = Accessibilism): Internalism is the view that, if two people are the same with respect to what it is rational for them to believe on the basis of introspection, then they are the same with respect to what it is rational for them to believe.

[14] A separate question is whether the state that ends up shaping the perception is itself introspectively accessible. Even if the state is accessible, that is not enough to remove the potential challenge to internalism, since the subsequent process leading to the perception might easily still fail to be introspectively accessible.

If the bad news about cognitive penetration of perception is true, two people can be the same with respect to what it is rational for them to believe on the basis of introspection, and yet not the same with respect to what it is rational for them to believe.
So,
If the bad news about cognitive penetration of perception is true, then internalism is false.

The problem is that the equation of internalism with accessibilism is entirely wrong. The two views are not equivalent, and accessibilism is not even a statement of a flavor of internalism in epistemology.

To see the main problem, notice how accessibilism is formulated in terms of the notion of what it is rational for us to believe on the basis of introspection, while not saying anything about what it takes for something to be rational for us to believe on the basis of introspection. Now, we should not assume that some internalist picture of introspection is correct---we need to have room for internalist/externalist disputes about introspection itself. For example, there is room for debate about whether an evil demon could make us radically mistaken about our own minds, and thereby impede our ability to form rational beliefs about our minds on the basis of introspection. However, if we try capture such a debate along the line of accessibilism, we produce only the following triviality:

If two people are the same with respect to what it is rational for them to believe on the basis of introspection, then they are the same with respect to what it is rational for them to believe on the basis of introspection.

There is no room for debate about the above, but there is room for an internalism/externalism debate about introspection. So there must be some way other than accessibilism to articulate what it is in question in internalist/externalist disputes.[15]

[15] My point relies on the epistemic characterization of access in accessibilism, and Lu Teng pointed out to me that there is room for non-epistemic psychological formulations of accessibilism, say in terms of mental states that are registered by inner perception or some other form of internal monitoring mechanism. In response, I would say that the move collapses the view into a form of mentalism, the sort of internalism that formulated

I won't now try to reach a better understanding of internalist/externalist disputes in epistemology (having already tried to do so in my 2018). Here I want to emphasize that, even if I'm wrong about the upshots of bad cognitive penetration of perception, this would not yet favor the rejection of internalism in epistemology.

There is an objection I should address. You might object that there's no substantive issue about how to formulate internalism in epistemology, and say that it doesn't matter whether accessibilism fails be a form of internalism or not. Perhaps what matters is whether accessibilism is true, not whether it is a version of internalism.

We can respond to the objection by shifting our main point, and setting the question of what counts as internalism aside. Given that accessibilism uses the notion of rationality in its antecedent, it fails be a theory of what it takes to have a rational belief, and in particular does not tell us anything about what it takes for it to be rational to believe something through introspection. Given that we should be seeking a broader theory about what it is rational to believe something, and accessibilism does not provide such a theory, the interest of accessibilism is drastically diminished. Versions of internalism, as broad theories of what it takes for it to be rational to believe something, are instead the claims to be interested in here. They have yet to be challenged by cases of cognitive penetration. Even if it turned out that the rational status of our beliefs is undermined by the inner workings of our minds, internalists could take that fact on board.

Conclusion

There's ever so much bad news these days, including bad news about the inner workings of our minds. According to many contemporary philosophers, this bad news has negative implications for the rationality of our beliefs drawn from such central sources such as moral intuition, introspection, and perception. I have argued against the application of this line of thought to perception, by zooming in on

---

in terms of types of mental states rather than forms of access to them (see e.g. Wedgwood 2000, forthcoming for more on mentalism vs. accessibilism).

what exactly happens if we learn bad news from psychology about our minds. Since the rational status of the relevant belief seems to be defeated as a result of learning the bad news, the truth of the bad news about the perception must have been compatible with the positive rational standing of the relevant belief after all.

While I have framed my central line of argument in section 3 in negative terms, as an objection to the view of bad cognitive penetration by figures such as Markie, Siegel, and McGrath, I in effect have argued for a positive claim about our central example of bad cognitive penetration. When John's perception of the face as angry is generated by his racial bias or prejudice, his perception does not fail to be a source of any rational support for his belief, and therefore actually is a source of some rational support for his belief. Given that a racial bias is driving John's perception, our result is troubling.

It is important to see that my positive conclusion here is qualified. It leaves open whether we could ever get all the way to knowledge on the basis of perception when our perception is the result of bad cognitive penetration. And it is silent on what sort of bar John's belief would have to meet in order to make subsequent action by John rational. Our result remains troubling all the same. Here I think it is most important to stress that we should not assume the domain of rationality to somehow be immune from the impact of racial bias or prejudice.[16] Such an assumption about the domain of rationality would be highly utopian. We have seen even more reason to believe that we are not in any utopia.

References

Alston, W. P. (1980). Level-confusions in epistemology. *Midwest Studies in Philosophy*, *5*(1), 135-150.

Basu, Rima (2019). The wrongs of racist beliefs. *Philosophical Studies* 176 (9):2497-2515.

[16] For relevant further discussion, see Haslanger 2007, Gendler 2011, Basu 2019, or Srinivasan forthcoming.

Bayne, T., & Spener, M. (2010). Introspective humility. *Philosophical Issues*, *20*, 1-22.
Burge, T. (2003). Perceptual entitlement. *Philosophy and phenomenological research*, *67*(3), 503-548.

Bijlstra, G., Holland, R. W., and Wigboldus, D. H. J. (2010). The social face of emotion recognition: evaluations versus stereotypes. *J. Exp. Soc. Psychol*. 46, 657–663.

Brownstein, M. 2018. *The Implicit Mind: Cognitive Architecture, the Self, and Ethics*. Oxford University Press.

Carruthers, P. (2011). *The opacity of mind: an integrative theory of self-knowledge*. OUP Oxford.
Christensen, D. (2010). Higher-Order Evidence 1. *Philosophy and Phenomenological Research*, *81*(1), 185-215.

Churchland, P. (1988). "Perceptual Plasticity and Theoretical Neutrality: A Reply to Fodor" *Philosophy of Science* 55, 167-87.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of personality and social psychology*, *108*(2), 219.

Jeffrey Moran and Robert Desimone, "Selective Attention Gates Visual Processing in the Extrastriate Cortex," Science, ccxxix, 4715 (Aug. 23, 1985): 782–84, at p. 783.

Doris, J. M. (2015). *Talking to our selves: Reflection, ignorance, and agency*. OUP Oxford.

Firestone, C., & Scholl, B. J. (2015). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, 1-72.

Fodor, J. (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press

Gendler, T. (2011). On the epistemic costs of implicit bias. *Philosophical Studies* 156(1):33-63.

Francis Gregory. (2016). Implications of "Too Good to Be True" for Replication, Theoretical Claims, and Experimental Design: An Example Using Prominent Studies of Racial Bias. *Frontiers in psychology*, *7*, 1382. https://doi.org/10.3389/fpsyg.2016.01382

Hansen, T. Olkkonen, M., Walter, S. & Gegenfurtner, K. R., (2006). "Memory Modulates Color Appearance". *Nature Neuroscience*, 9(11), 1367–1368.

Haslanger, Sally (2007). "But mom, crop-tops are cute!" Social knowledge, social structure and ideology critique. *Philosophical Issues* 17 (1):70–91.

Hugenberg, K. (2005). Social categorization and the perception of facial affect: Target race moderates the response latency advantage for happy faces. *Emotion, 5,* 267–276. http://dx.doi.org/10.1037/1528-3542.5.3.267

Hutchings, P. B., & Haddock, G. (2008). Look Black in anger: The role of implicit prejudice in the categorization and perceived emotional inten- sity of racially ambiguous faces. *Journal of Experimental Social Psy- chology, 44,* 1418–1420. http://dx.doi.org/10.1016/j.jesp.2008.05.002

Kotzen, M. (2019). A formal account of epistemic defeat. In *Themes from Klein* (pp. 213-234). Springer, Cham.
Lasonen-Aarnio, M. (2014). Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, *88*(2), 314-345.

Ling, Sam, Taosheng Liu, and Marisa Carrasco. 2009. "How spatial and feature-based attention affect the gain and tuning of population responses." *Vision Research* 49, no. 10: 1194-1204.

Jack Lyons (2011). Circularity, Reliability, and the Cognitive Penetrability of Perception. Philosophical Issues 21 (1):289-311.

Machery, Edouard (forthcoming). The Alpha War. *Review of Philosophy and Psychology*:1-25.

Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, *84*(1), 24-62.

Mandelbaum, Eric. "Attitude, inference, association: On the propositional structure of implicit bias." *Noûs* 50.3 (2016): 629-658.

Schoenfield, M. (2018). An accuracy based approach to higher order evidence. *Philosophy and Phenomenological Research*, *96*(3), 690-715.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review*, *117*(2), 245-273.

----- (2011). *Perplexities of consciousness*. MIT press.
Susanna Siegel (2013). The epistemic impact of the etiology of experience. *Philosophical Studies* 162 (3): 697-722.
Siegel, S. (2017). *The rationality of perception*. Oxford University Press
Siegel, Susanna (forthcoming). Bias and Perception. In Erin Beeghly & Alex Madva (eds.), *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*. Routledge.
Silins, N. (2016). Cognitive penetration and the epistemology of perception. *Philosophy Compass*, *11*(1), 24-42.

Sinnott-Armstrong, W. (2006) Moral intuitionism meets empirical psychology in *Metaethics after Moore* (ed) Horgan, T. Oxford, UK: Oxford University Press.
Smithies, Declan (2019). *The Epistemic Role of Consciousness*. New York, USA: Oxford University Press.

Srinivasan, A. forthcoming. Radical Externalism.

Stokes, D. (2012). Perceiving and desiring: A new look at the cognitive penetrability of experience. *Philosophical studies*, *158*(3), 477-492.
Vance, Jonna (2014). Emotion and the new epistemic challenge from cognitive penetrability. *Philosophical Studies* 169 (2):257-283.

Weatherson, B. (2019). *Normative externalism*. Oxford University Press.

Wedgwood, Ralph (2002). Internalism Explained. Philosophy and Phenomenological Research 65(2): 349– 369.
(forthcoming). Internalism Re-explained. In Julien Dutant (ed.), The New Evil Demon: New Essays on Knowledge, Justification and Rationality . Oxford: Oxford University Press. Available at http://www-bcf.usc.edu/~wedgwood/Internalism_reexplained.pdf

White, R. (2010). You just believe that because... *Philosophical Perspectives*, *24*, 573-615

Wittenbrink B., Correll J., Ma D.S. (2019) Implicit Prejudice. In: Sassenberg K., Vliek M. (eds) Social Psychology in Action. Springer, Cham