The Philosophical Quarterly Vol. 0, No. 0 *2024*
ISSN 0031-8094 https://doi.org/10.1093/pq/pqae117

Emotional unreliability and epistemic defeat

Tricia Magalotti

Stockholm University, Sweden

Among those who think that emotions can provide epistemic reasons for belief, there is disagreement about whether emotions provide foundational reasons (ones that are not based on further reasons) or non-foundational reasons (ones that are based on further reasons). I argue in this paper that considerations about evidence of emotional unreliability favour the non-foundational view of emotional reasons. The argument starts with a set of counterexamples to the claim that evidence of emotional unreliability always defeats emotional justification. I then show why only the non-foundational picture of emotional reasons is compatible with this finding. The upshot is 2-fold: first, the commonly held assumption that evidence of emotional unreliability always defeats emotional justification is false; and, second, this gives us a reason for preferring a non-foundational picture of emotional reasons.

Keywords: Emotion; Epistemology of emotion; Perceptualism; Defeaters; Emotional dogmatism; Emotional justification.

I. Introduction

It is widely assumed, among philosophers and non-philosophers alike, that if we have evidence that (some of) our emotions are unreliable, then this would prevent (those) emotions from providing epistemic reasons for belief. The idea is that even if emotions were capable of epistemically justifying belief, this justification would be defeated by evidence of unreliability. I argue here that this assumption is false. It is not always the case that evidence of emotional unreliability defeats emotional justification. This is the first aim of the paper.

My second aim is to show that if there are counterexamples to the claim that evidence of emotional unreliability always defeats emotional

Correspondence to: Tricia Magalotti, tmagalotti@gmail.com

justification, then this presents a problem for theories of emotion according to which emotions provide foundational reasons for belief. Such theories must treat evidence of emotional unreliability as undercutting defeaters. I will also argue that theories of emotion according to which emotions provide non-foundational reasons for belief are compatible with the denial of the claim that evidence of emotional unreliability always defeats emotional justification. These theories can treat evidence of emotional unreliability as higher order evidence. And on at least some views of higher order evidence of unreliability, such evidence does not always defeat one's justification for the relevant mental states.

The plan for the paper is this. In Section II, I discuss the different views under consideration and motivate the more general position that emotions can provide epistemic reasons for belief. In Section III, I offer three counterexamples to the claim that evidence of emotional unreliability always defeats emotional justification. In Section IV, I explain why this is problematic for the foundational view of emotional reasons. And in Section V, I explain why the view that emotions provide non-foundational reasons for belief can accommodate the counterexamples from Section III.

II. Emotions as reasons for belief

There are three available views regarding the justificatory relationship between emotions and belief. *Non-rationalism* holds that emotions do not provide epistemic reasons for belief. Emotions may perhaps focus our attention towards important epistemic reasons, but one cannot rationally base beliefs on emotions (Brady 2013). According to *rationalism*, emotions do provide epistemic reasons for belief. It has two versions, perceptualism and inferentialism.

Perceptualism holds that emotions provide *foundational* reasons for belief, reasons that are not themselves based on any further reasons (Elgin 1996; Döring 2008; Tappolet 2016; Mitchell 2021).[1] For perceptualists, emotions are epistemically symmetrical to perceptions. Just as perceptual appearances provide foundational reasons for empirical beliefs, so too emotional 'appearances' provide foundational reasons for evaluative beliefs. While emotional perceptualists often endorse *emotional dogmatism*, the view that emotions provide immediate prima facie justification for evaluative beliefs, they need not. Just as there are dogmatists and non-dogmatists about perceptual justification, so too emotional perceptualists can be dogmatists or non-dogmatists.

[1] I am concerned with *epistemic* perceptualism rather than *representational* perceptualism, according to which the representational content of emotions is similar to that of perception (Tappolet 2021).

The third view, which I will call *inferentialism*, holds that emotions provide *non-foundational* epistemic reasons for belief (Solomon 1977; Nussbaum 2001; Deonna and Teroni 2012).[2] Inferentialists agree with perceptualists that emotions provide reasons for belief. But, unlike perceptualists, they think that emotions are based on further reasons. Emotional inferentialism is typically associated with judgmentalist views of emotion, according to which emotions are (or are relevantly similar to) evaluative judgments. But inferentialists are not committed to judgmentalism. It could be, for example, that emotions are a *sui generis* mental state type and yet, like judgments, they are epistemically based on evidential reasons. If inferentialism is true, then it seems that emotions will themselves be subject to epistemic evaluation. Emotions can be either rational or irrational depending on whether the reasons they are based on are good or bad.

We can think of emotional reasoning as a two-step process. The first step begins with some 'emotional antecedents' (as they are called in the psychology literature) that cause the emotion. For example, if one sees a dog and becomes afraid of the dog in response, the visual experience of the dog is the emotional antecedent. I will call this causal link between the emotional antecedents and the emotion itself *Link 1*. For perceptualists, Link 1 is merely causal. Just as the causal processes that give rise to a visual experience are a-rational, so too are the causal processes that give rise to emotional experience. For inferentialists, on the other hand, Link 1 is also a reasoning link. The emotional antecedents, such as the visual experience of the dog, not only *cause* the resulting emotion but also, at least in the good case, *rationalize* it. When the emotion is based on good reasons, the emotion is epistemically justified or rational. In the second step, which I will call *Link 2*, the subject treats the emotion as an evidential reason for belief. Fear of the dog provides a reason for believing that the dog is dangerous. Both perceptualists and inferentialists treat Link 2 as a reasoning link.

The question I wish to address here is whether evidence of emotional unreliability always defeats emotional justification. This is a crucial question for all rationalists, because if evidence of unreliability did always defeat justification, this would be a potential threat to rationalism since evidence of emotional unreliability appears to be abundant.[3] I will argue that evidence of emotional unreliability does not always defeat emotional justification. One might think that this is all the better for both versions of rationalism, but I will argue that ultimately it presents a problem for perceptualists.

There are two assumptions worth addressing at the outset. First is rationalism itself. My assumption of rationalism here is mainly dialectically

[2]Solomon and Nussbaum are judgmentalists. Deonna and Teroni argue for a distinct view called 'attitudinalism'.

[3]Carter (2020) calls this the 'reliability problem'.

motivated. First, I am considering a thesis that is a potential challenge to rationalists, the threat of evidence of emotional unreliability to emotional justification. Secondly, my rejection of the claim that evidence of emotional unreliability always defeats emotional justification will be leveraged in order to settle an in-house dispute among rationalists, namely whether emotional reasons are foundational or non-foundational. In service of these aims, I will assume that emotions can in fact provide epistemic reasons for belief. But I do think it is worth saying a few words about why one might be motivated to accept this.

To this end, I want to discuss two examples. The first is Elgin's (1996) case of Joan, who feels envious of her colleague Felix:

> Joan's envy over Felix's promotion to the Mugwump Chair of Metaphysics heightens her awareness of disparities in the reception of their respective views. It brings into focus the widespread practice of casting women's professional activities in a less favourable light than men's. Felix's manner is considered self-assured; Joan's is said to be aggressive. His first draft shows promise; hers needs work. His criticisms are incisive; hers are carping. His work extends the insights of his predecessors; hers is derivative. The list seems endless. (Elgin 1996: 149–50)

Elgin focuses on the positive epistemic consequences of Joan's envy, which allows her to process a body of evidence in a way that she would not have otherwise been psychologically poised to do. But, things can also go the other way around. Emotions can be *responsive to* certain patterns in the evidence that we would not have noticed otherwise. For example, sometimes our anger picks up on and alerts us to certain injustices that otherwise might have passed us by. Suppose that, rather than feeling jealous, Joan felt angry that Felix had been promoted. This anger, let's say, was sensitive to all of the same instances of differential treatment of women and men Elgin lists. Joan may be unable to identify the relevant pattern upon reflection and, if she had lived in a certain time and place, may have even lacked the conceptual resources to form the relevant beliefs. Emotions can clue us in to certain non-obvious implications of our evidence. In such cases, it seems appropriate for one to treat one's emotion as a reason for believing, for example, that one's being passed over for a promotion was unfair.

The second example, which is discussed in a different context by Arpaly (2003), is the case of Huck Finn from Mark Twain's (1884/1994) *The Adventures of Huckleberry Finn*.[4] In the novel, Huck finds himself with Jim, who has recently escaped slavery in pre-Civil-War Missouri. Huck believes that the morally right thing to do is turn Jim in to Miss Watson, Jim's purported 'owner'. Huck writes a letter, intending to send it to Miss Watson for this purpose. However, as Huck builds an emotional connection with Jim, an inner conflict starts to brew. It is important to note that Huck is not wondering what is the *right* thing

[4]The example is also used earlier by Bennett (1974).

to do. He firmly believes that the morally right thing to do is to turn Jim in. It is rather that he struggles with whether or not to do 'the right thing'. In the end, Huck tears up the letter in an act of apparent moral akrasia and proceeds to help Jim escape.

Since Huck does not tear up the letter because he thinks that this is what he morally *ought* to do, we need some other explanation for why he acts as he does. Most plausibly, what motivates Huck, in large part, is the emotion that he feels towards Jim, some variety of compassion that is responsive to features of Jim's humanity that become evident to Huck as they cultivate a relationship. But the compassion that Huck has for Jim is not merely a motivating reason. It also seems to go some way towards *rationalizing* or *justifying* Huck's action of tearing up the letter. The compassion, then, in addition to being a motivating reason for belief, is also a normative reason.[5] If Huck's compassion can be a normative reason for *action*, then it seems plausible that it could also be a normative reason for *belief*. For example, in addition to being motivated to help Jim, Huck's compassion might cause him to form the belief that Jim's interests matter, or, less intellectually, that what happens to Jim matters (for Jim's sake). It seems here that the compassion is not merely an a-rational cause of the belief. It provides some justification for that belief.

One of the main motivations underlying rationalism is this thought that emotions psychologically mobilize certain epistemic reasons that the subject already possessed but was not in a position to draw the relevant conclusions from. This means that they play an important casual role in the production of belief. However, there is some intuitive sense in which the emotion not only plays a *causal* role but also provides some *epistemic justification* for the belief. Joan's anger seems to provide some justification for the claim that she has been treated unfairly. And Huck's compassion seems to provide some justification for thinking that what happens to Jim matters.

Notice that the descriptions of these cases are compatible with both perceptualism and inferentialism. I have stipulated that Joan's anger and Huck's compassion are sensitive to their background beliefs. One might worry that this rings inferentialist. But this worry is misplaced. Perceptualists, after all, do not deny that emotions can be caused by beliefs. Nor are they committed to denying that emotions can be sensitive to certain evidential properties of a body of beliefs. Rather, they are committed to holding that even when emotions are caused in this way, they are not *rationally based* on those beliefs.

The second important assumption I will make is that of a broadly evidentialist framework according to which epistemic reasons for belief are evidential reasons. Accordingly, I will use the terms 'evidence' and 'epistemic reason for belief' more or less interchangeably. One could also give a non-evidentialist analysis of the epistemology of emotion. Carter (2020), for example, argues

[5]See also Tappolet (2016) and Silva (2022) on emotion-based inverse akrasia.

for a virtue reliabilist account of emotional justification according to which emotions can justify beliefs when both the emotion and the resulting belief are formed in a reliable way, which for Carter involves the manifestation of certain 'competences. There are two reasons why I adopt the evidentialist framework. First, talk of reasons and evidence is highly intuitive, and, partly as a consequence of this, I think that the question of whether and how emotions can *provide good evidential reasons* for belief is an interesting one worthy considering in its own right, independently of whether emotions can figure in a reliable process of belief formation. And it is unclear how to make sense of reasons within a reliabilist framework. Secondly, one of the main questions that this paper is meant to answer is whether evidence of unreliability defeats emotional justification. Again, there is no straightforward treatment of this question from the reliabilist perspective since what matters from that perspective, in the first instance, is *actual* unreliability rather than *evidence* of unreliability. Nonetheless, for those who are interested, I have included some brief remarks at the end of Section V regarding the impact of the cases I discuss for reliabilist treatments of emotional reasoning.

To sum up, both perceptualism and inferentialism are versions of emotional rationalism, the view that emotions can serve as epistemic reasons for belief. Emotional rationalism is motivated by the observation that there seem to be cases where emotions do provide epistemic reasons to accept evaluative claims. In the next section, I will argue that it is plausible that emotions can provide epistemic reasons even when the subject has evidence that the emotion type is unreliable. Following that, I will explain why only inferentialists, and not perceptualists, are in a position to accommodate this observation.

III. Emotional justification despite evidence of emotional unreliability

The prevailing opinion is that evidence that one's emotions are unreliable defeats the emotional justification for belief that otherwise would have been provided. If one has evidence that one's emotions don't reliably get things right, then one should not rely on one's emotions as epistemic reasons. This is the implicit assumption in slogans such as 'don't trust your feelings, they lie'. In this section, I argue the prevailing view is false.

A few points of clarification are in order. First, unreliability in this context should be interpreted as *epistemic* unreliability. In order for a process type to be reliable, it must have a tendency to produce true rather than false beliefs (or, in this case, correct rather than incorrect emotions). There are many further details to work out (e.g. actual vs hypothetical reliability, which ratio of true to false beliefs is required to meet the relevant threshold, etc.), but we need not settle these finer details here. The claim that emotions are unreliable, then, is

to be interpreted in the following way: the psychological processes that produce emotions do not produce a high ratio of correct to incorrect emotions. If one realizes that one is too easily angered for example, that one feels anger in many situations where the anger gets things wrong, then one has evidence that one's anger is unreliable in the relevant sense.

Relatedly, I want to emphasize that the unreliability that I am concerned with here is Link 1 unreliability, unreliability in the production of the emotion itself. This is not the only possible locus of unreliability in the emotional reasoning process. Indeed, emotional reasoning processes have also acquired a reputation for their purported Link 2 unreliability. Emotions, the thought goes, cause us to reason in irrational ways and, thus, to form an unacceptable ratio of false to true beliefs.[6] The reason that I focus here on Link 1 unreliability is that Link 1 is where we find the epistemologically relevant difference between perceptualism and inferentialism. And it is their divergent treatment of this link, I will argue, that is the source of an important advantage of inferentialism.

Next, unreliability claims can be levelled at a global or local level. A *global* unreliability claim targets all emotions: emotions, in general, are unreliable. A *local* unreliability claim targets some subclass of emotions. This could include either specific emotional attitude types (e.g. anger or fear), or the emotions of a specific subject (e.g. Mary's emotions rather than Paul's). If a subject receives evidence that her emotions of type T are unreliable, then it is the justification provided by T-emotions that is purportedly defeated by the evidence of unreliability. Sometimes I will use the shorthand 'evidence of emotional unreliability does (not) defeat emotional justification', but this should always be understood to include implicit reference to the relevant subclass of emotions.

My argument in this section rests on three counterexamples to the universal claim that evidence of emotional unreliability always defeats emotional justification. The cases are ones in which the subject has evidence that their emotions of a certain type are unreliable, and yet it seems that the subject is epistemically permitted to use an emotion of that type as a reason for belief. If it is true that *any* of these subjects is rationally permitted to use their emotion as a reason, we have a counterexample to the claim that evidence of emotional unreliability always defeats emotional justification.

Helena the Hothead

Helena is overly prone towards anger. She gets angry at the bus driver for showing up two minutes late in the middle of a snowstorm, at her two-year-old for accidentally spilling his milk, and at elderly people who move slowly in the supermarket. Helena knows that she is a bit of a hothead and that

[6] As Goldie says, they 'skew the epistemic landscape' (Goldie 2008).

her anger is not a reliable indicator that someone has in fact wronged her or anyone else.

Helena, as it happens, is part of the same department where Felix has just been chosen over Joan to be promoted to the Mugwump Chair of Metaphysics. She, like Joan, notices a string of individual instances of differential treatment of men and women in the department, and in particular, of Joan and Felix. Additionally, Helena finds Joan's philosophical work to be superior to Felix's. And she sees Joan's many contributions to the department in terms of service as well as her dedication to her students. When Helena hears that Felix was promoted rather than Joan, she feels angry about it (on Joan's behalf). Her anger is caused by all of this background information, information that, at least to the right eyes, exhibits a distinctive evidential pattern. Let us suppose that if this pattern had not been there, Helena would not have felt angry in these particular circumstances. Furthermore, if Helena had never experienced the anger, she would not have been in a position to integrate all of that information in the right way to form the judgment that Helena had been wronged. Even after experiencing the emotion, if Helena were to reflect about whether or not Joan was wronged by being passed over, she might not be able to formulate all of these background beliefs into a cohesive argument in favour of that conclusion. It seems to me that Helena's unreliability with respect to anger is not enough to rationally disqualify her anger from counting as a reason. To the contrary, Helena's anger—in this case, though not in general—serves as an important source of insight. So, it might be rational for her to use her anger as a reason for believing that Joan was wronged or as a reason to believe that she should tell Joan that she thinks that the decision was unfair. But if evidence of emotional unreliability always defeats emotional justification, then Helena is rationally prohibited from treating her anger as a reason, even if it is responsive to patterns of information in her background beliefs.

Hans the Hard Determinist

Let us consider one final colleague of Joan, Felix, and Helena. Hans, who happens to be the head of the department, is also a committed hard-determinist. He believes that we do not have free will and, as a result, that no one is ever blameworthy for their actions. Let's suppose that, whether or not this belief is true, it is justified for Hans. He also believes that the formal object of moral indignation is moral blameworthiness and, thus, that moral indignation is never an appropriate emotion to feel (in the sense that it never gets things right about how the world actually is). Despite these theoretical commitments, Hans nonetheless finds himself feeling indignant fairly regularly and has a justified belief that his indignation responses are unreliable.

After Felix's promotion, it comes to Hans's attention that Felix had covertly promised the members of the selection committee a trip to his summer home

in Spain. Despite his commitment to hard determinism, he finds himself feeling indignant about the committee's choice to promote Felix. Hans tries to suppress this emotion by reminding himself that his indignation is unreliable. But, it seems that Hans is epistemically permitted to treat his indignation as a reason despite this evidence of unreliability. It is a consideration in favour of believing that the committee members *deserved* some sort of sanction. If evidence of emotional unreliability always defeats emotional justification, then the epistemic justification that his indignation provides for that belief should be defeated, leaving the belief unjustified. But, this seems not to be the case.

Chuck Finn

Finally, consider the case of Chuck Finn, Huck Finn's overly compassionate counterpart. Chuck is in the same position as Huck with the exception that he feels compassion towards inappropriate targets such as books, tables, mugs, and lamps. When someone accidentally knocks a mug off the table, he feels compassion for the mug as it tumbles across the floor. Because his family and friends chastise him for this, Chuck is well aware that his compassion is epistemically unreliable.

When Chuck Finn finds himself holding the letter to Miss Watson and debating about what to do, his compassion for Jim, like Huck's, is activated. Chuck's compassion, in this particular case, is caused by observations that he makes about Jim. Like Huck, Chuck is inclined, on the basis of this compassion, to believe that what happens to Jim matters and that he ought to rip up the letter. Once again, Chuck has reasons that support the conclusion that what happens to Jim matters. But, without the emotion, these reasons remain psychologically inert. Of course, Chuck has evidence that his compassion is unreliable. But, this does not appear to rationally disqualify Chuck's compassion as a reason. If Huck is epistemically permitted to use his compassion as a reason, it seems that Chuck is epistemically permitted to use his. So, here too we have a case where evidence of emotional unreliability does not seem to defeat emotional justification to believe.

As explained above in Section II, it is perfectly compatible with perceptualism to claim that emotions are caused by beliefs and even that emotions can be sensitive to evidential properties of the subject's background beliefs. Indeed, this is in line with the basic motivation for rationalism generally. So, my stipulating that the emotions of the various protagonists are sensitive in this way is not question-begging against the perceptualist. Perhaps the perceptualist would wish to deny that, as a matter of empirical fact, emotions do respond to evidential properties of bodies of beliefs. However, this strikes me once again as being out of touch with the basic motivation for emotional rationalism. Emotions seem to provide us with a certain kind of epistemic insight. They allow us to see certain evaluative conclusions that otherwise would

have remained psychologically opaque. But in order to provide us with this insight, rather than just a random assortment of information, it seems that emotions must, to some degree, be sensitive to the evidential properties of the mental states that are their causal antecedents. One can deny that this sensitivity is sufficient for inference, but to deny the sensitivity itself runs counter to the basic rationalist picture.

One might wonder whether it would not be better, in response to these cases, to go non-rationalist. If Helena, Hans, and Chuck retain justification in the face of emotional unreliability, then perhaps we should think that it was not their emotion doing their justificatory work in the first place. While this is one way of responding to the cases, I think that there are reasons for resisting it.

For some of us, it seems intuitive that our emotions do not just *cause* beliefs but, at least in some circumstances, are capable of *rationalizing* those beliefs. This is more or less what I tried to show in discussing the initial motivations for rationalism in Section II. But to elaborate on this point a bit, it seems to be the case that Chuck's compassion for Jim is not only causally relevant to his decision not to turn Jim in. It is *telling* him something about how things are, and it is doing so in a rational way. There are cases in which the subject is unable to cite the mental states that are the causal antecedents of their emotions as reasons and yet are able to identify the emotion itself as something that rationally counts in favour of the relevant action or belief. This seems to be the case for both Helena and Chuck. In these cases, it seems correct for the subject to cite their emotion as a reason for, a consideration in favour of, an action or belief. This is especially important in cases of conscious deliberation. Not all cases of action or belief formation are conscious in this way, but we do want to be able to accommodate the ones that are. The non-rationalist, however, must say that if Chuck can cite *only* his emotion and not its causal antecedents, then he is incapable of identifying any of his reasons for belief or action. To my mind, this is an undesirable result.

For those who are moved by these sorts of considerations, I take it that they will feel some pull towards countenancing the reason-providing potential of the relevant emotions in the cases of Helena, Hans, and Chuck. The bigger picture is that emotions serve as important sources of insight into the evaluative realm. What the above examples are meant to highlight is that evidence of an emotion type's unreliability (for a subject) does not always preclude it from playing that insight-providing role. To insist that subjects such as Helena, Hans, and Chuck are rationally disallowed from treating their emotions (of the relevant types) as epistemic reasons would run contrary to this basic conviction underlying emotional rationalism. To the extent that one feels that pull, one should prefer a view that allows for the possibility that some emotions can justify even when the subject possesses evidence of emotional unreliability.

Finally, before moving on, I return briefly to the issue of what impact these cases have for reliabilist views of emotion. Foundationalist reliabilists, like foundationalist evidentialists, are concerned about whether emotions are foundational or non-foundational elements of the justificatory chain. Ideally, one would be able to present an argument that gives a parallel treatment of this issue in both frameworks. Unfortunately, while the considerations that I have provided here are likely to have interesting implications for reliabilists, they are not relevant for reliabilists in the same way.

The primary problem is that, at least on the face of things, it seems that the reliabilist must deny that Helena and Chuck's emotions are justification-providing in the first place. This is because, in addition to having *evidence* of Link 1 unreliability, their emotional processes *actually* exhibit Link 1 unreliability. And if the emotion is unreliably produced, then it cannot confer even *prima facie* justification to the resulting belief. And so, there is no initial justification to be defeated. Their emotions fail to provide justification from the start.[7] If, to the contrary, their emotions are justification-providing (or, as I have been saying, 'reason-providing'), then this is a strike against a reliabilist analysis. Of course, there are things that a reliabilist might say in response to this sort of challenge. But which responses are available will depend on the particular brand of reliabilism, including their preferred way of individuating process types and what kind of reliability they require. Given the complexity of these issues, I will leave the implications for reliabilism for another time.

IV. Against emotional perceptualism

In the previous section I argued that there are cases in which emotional justification seems not to be defeated by evidence of emotional unreliability. In this section, I will argue that emotional perceptualists are committed to claiming that evidence of emotional unreliability is undercutting evidence, and therefore, that evidence of emotional unreliability does always defeat emotional justification. The conjunction of these two claims, of course, spells trouble for epistemic perceptualists.

Traditionally, epistemologists distinguish between undercutting and rebutting defeaters. More recently, there has also been extensive debate regarding whether and in what way higher order evidence can defeat one's evidential support for a belief. Having a grasp on these distinctions is essential for understanding why inferentialism has an advantage over perceptualism. I will review each of these three kinds of defeat in turn.

[7]This is true for both process reliabilism and virtue reliabilism. In Carter's (2020) virtue-style account, for example, neither Chuck nor Helena's emotions manifest the right kind of competence.

Rebutting defeaters are 'reasons for denying the conclusion', or reasons that attack the conclusion directly (Pollock 1987: 485). If one has a belief, formed on the basis of a lifetime of observational data, that all swans are white and then encounters a black swan, then new evidence that there is a black swan directly attacks the conclusion that all swans are white by providing a counterexample. *Undercutting defeaters* are ones that 'attack the connection between the reason and the conclusion' (*ibid.*). Perceptual defeat is a classic case of undercutting defeat. Suppose I believe that the wall in front of me is red on the basis of my visual experience. If I then learn that someone has pointed a red light onto the wall, then this new evidence undercuts the evidential relation between my perceptual experience and the belief that the wall is red. One way of understanding undercutting defeat is that it lowers the conditional probability of the conclusion on one's evidence. The conditional probability that the wall is red given that it looks red is high, high enough that its looking red makes it likely that it is red. But the conditional probability that the wall is red given that it looks red and that there is a red light shining on it is much lower, low enough that the evidence no longer makes it likely that the wall is red. And this lowering of the conditional probability seems to show how rebutting defeaters defeat.

Higher order defeat is more complicated. There is no consensus about whether higher-order evidence (HOE) even does defeat and, if so, how (more on this later). There are different characterizations of HOE, but I will appeal to one from Lasonen-Aarnio, who says that HOE (of the potentially worrying sort) is evidence that 'I am subject to a cognitive malfunction of some sort and hence that my doxastic state is the output of a flawed cognitive process' (2014: 315–16).[8] For short, we can say that HOE indicates that a belief was 'badly produced'. A classic example of HOE is a potentially hypoxic airplane pilot. Here is Horowitz's version:

> Aisha is out flying her small, unpressurized airplane, wondering whether she has enough fuel to make it to Hawaii. She looks at the gauges, dials, and maps, and obtains some evidence, E, which she knows strongly supports (say to degree .99) either the proposition that she has enough gas (G) or that she does not (∼G). Thinking it over and performing the necessary calculations, Aisha concludes G; in fact, this is what E supports. But then she checks her altitude and notices that she's at great risk for hypoxia, a condition which impairs one's reasoning while leaving the reasoner feeling perfectly cogent and clear-headed. Aisha knows that at this altitude, pilots performing the kinds of calculations she just did only reach the correct conclusion 50% of the time. (Horowitz 2022)

There are several characteristic features of HOE that seem to distinguish it from undercutting defeaters.[6] One that is especially helpful for present purposes is that HOE does not seem to lower the conditional probability of the

[8]Here I consider only negative HOE, since this is the kind relevant to questions of unreliability and higher order defeat.

conclusion on the evidence in the same way as undercutting defeaters do. This is easiest to see when we consider a case of conclusive support. Suppose, for example, Aisha's first order evidence (FOE), which she gets from the 'gauges, dials, and maps' and maybe some additional background evidence, *entails* the conclusion that she has enough gas. So, the conditional probability that she has enough gas given her FOE is 1. It follows that the conditional probability that she has enough gas given her FOE *and that she is hypoxic* is still 1. In the case of entailing FOE, then, HOE does not lower the conditional probability of the conclusion on one's evidence. Still, some think, Aisha is required to lower her confidence in the conclusion that she has enough gas to make it to her destination. So, HOE might defeat justification, but if it does, it seems to do so in a different way than ordinary undercutting defeat.[9]

With these distinctions in hand, we can return to the analysis of evidence of emotional unreliability. The unreliability is located in Link 1. For perceptualists, Link 1 is a mere causal link, not a reasoning link. Emotions are not *based* on evidence, so the process that produces the emotion cannot be construed as a reasoning process. Because of this, perceptualists are, unsurprisingly, required to treat evidence of emotional unreliability in the same way that we treat evidence of perceptual unreliability, namely as an undercutting defeater. If someone has evidence that they are overly prone to anger, this is the emotional analogue of someone who knows that they are wearing rose-coloured glasses. Just as the glasses-wearer loses their perceptual justification for believing that things are red, the easily angered person loses their emotional justification for believing that people have wronged them.

The inferentialist, on the other hand, construes the link between the causal antecedents and the emotion as a *reasoning* link and the causal antecedents themselves as evidential reasons, not mere causes. The evidence on which the emotion is based can thus be treated as FOE. The justificatory chain for the belief that is formed on the basis of the emotion traces back to that FOE. The evidence of emotional unreliability, then, is best interpreted as evidence that one's reasoning process that produced the emotion in response to the FOE was flawed. This makes it HOE of a flawed reasoning process.

An example should help to see the difference more clearly. Suppose that Eva has strict standards for herself as a friend and believes that she is morally required to meet them. One day, Eva is up late working to meet an unexpected deadline and forgets that she promised to help her friends move the following morning. She accidentally sleeps in and fails to show up at the agreed upon time. Eva's background beliefs about what it takes to be a good friend and her belief about not showing up cause her to feel guilty. On the basis of this guilt,

[9]When it comes to non-entailing evidence, things are more complicated. But some think that even then, the conditional probability of the conclusion on the subject's evidence is not lowered (see also Christensen 2010 and Lasonen-Aarnio 2014).

she forms the belief that she is blameworthy for letting her friends down. Eva then calls her brother Max, who manages to present Eva with compelling reasons for thinking that her guilt responses are in general unreliable. She has a tendency to feel guilty in situations in which guilt is not appropriate, situations in which she has not acted wrongly. Max manages to convince Eva (rightly, we can suppose), and so she acquires evidence that her guilt is unreliable.

According to the perceptualist, Eva's background beliefs are mere causes, not evidence. When Eva gets evidence of the unreliability of her guilt response, then, this functions as undercutting evidence. One is only epistemically permitted to accept emotional 'appearances' (in this case, the appearance of one's being blameworthy) so long as one lacks evidence of unreliability. Since Eva has acquired evidence of the unreliability of her guilt responses, she is not epistemically permitted to treat that emotion as a reason any longer. In probabilistic terms, its (emotionally) appearing to Eva that she is blameworthy *and* the unreliability of her emotional appearances of blameworthiness together fail to make it likely that she is blameworthy. Thus, her emotional justification is undercut by the evidence of unreliability provided by Max. As a result, the evidential connection between the original evidence (her guilt) and the conclusion (the belief that she is blameworthy for failing to show up) is undercut and the justification is defeated. Link 2 is undermined as the result of acquiring evidence of unreliability in Link 1.

The inferentialist, on the other hand, claims that Eva's background beliefs are not only causes but also evidential *reasons* for the emotion. This is not to say that they are necessarily *good* reasons, but rather that Eva *treats* them as reasons. To be a reason in the sense, I mean here is a descriptive rather than an evaluative claim. They are the FOE for Eva's emotional conclusion, namely her guilt over leaving her friends in the lurch. When Eva receives evidence of emotional unreliability on this account, she receives evidence that the *reasoning process* that took her from her FOE to her emotional conclusion could have been faulty because she has an unreliable track record with reasoning processes of that type (viz guilt-producing processes). This is parallel to Aisha's situation when she learns that she could be hypoxic and, thus, that the reasoning link between her FOE and her conclusion about having enough gas could have been faulty. There is an objective fact about whether or not Eva's FOE (her beliefs about the moral standards for friendship and her failing to show up) supports the conclusion that Eva is blameworthy. The evidence supports that conclusion to some degree or other. Whether or not Eva's guilt response is reliable does not raise or lower this probability, because whether Eva feels guilty is not evidentially relevant to whether she is blameworthy. (Again, if one has a difficult time imagining this, one can suppose that the FOE is entailing.) Whether Eva's HOE defeats her belief that she is blameworthy is still left open here. But if it does, then this will be a case of higher order defeat rather than undercutting defeat.

One potential reply here is that perhaps the perceptualist could get around this problem by positing a degreed rather than binary notion of defeat according to which a defeater can *reduce* one's justification without lowering it below the threshold required for the belief to remain justified. On this line of defence, Helena's evidence that her anger is unreliable diminishes or downgrades her justification for believing that Felix's promotion over Joan was unjust, but it does not lower justification below the threshold necessary for the belief to count as justified. And that is why her emotion is still reason-providing.

The problem with the degrees of defeat response is that for it to work, Helena must retain some substantial justification for believing that the promotion was unjust. But it is hard to see on the perceptualist picture how she could retain the requisite level of justification. If the emotion itself is the ultimate source of justification, as the perceptualist argues, and Helena learns that her emotions of that type are unreliable, then what is the source of the justification that remains in the wake of the defeating evidence?

Of course, there are degrees of reliability (and the term 'reliable' is probably vague). If R is my perceptual reason for believing p, and I learn that perceptual processes of R's type are just a little bit unreliable (less than perfect), then I might retain justification for believing p. But, to be 'just a little bit unreliable' is, in fact, to be pretty reliable. Perfection is not required for reliability. It would be strange to use 'unreliable' to indicate the sort of process that is just slightly unreliable. To say that a process is unreliable is to say that it is pretty far off the mark. And it seems reasonable to say that when there is a singular ultimate source of justification for a belief, that unreliability in that source will entail defeat of justification in the binary sense. It will knock justification down below the threshold required for justified belief. If we are willing to apply the term 'unreliable' to a perceptual process type, then chances are that things are bad enough to undermine perceptual justification. If emotions are foundational reasons, then the same is true for emotional justification.

A second potential objection to my conclusion about the failure of perceptualism is that the sorts of cases I have discussed here involve relatively sophisticated emotions that are dependent on the background beliefs of the subjects. One might doubt that we would see similar effects in cases of less sophisticated emotions such as primal fear or disgust, and thus doubt that sophisticated and unsophisticated emotions can be treated in the same way. In particular, perhaps inferentialism gets it right about sophisticated emotions and perceptualism gets it right about unsophisticated emotions.

In addition to general (defeasible) considerations about parsimony, I think that there is some initial reason to resist this line. First, I do not find it obvious that the above cases, especially those of Helena and Chuck, involve particularly sophisticated emotions. Anger, for example, is typically classified

as one of the evolutionary and developmentally 'basic' emotions.[10] And the sort of compassion that Chuck feels for Jim seems to be of a piece with the sort of compassion that young children are capable of feeling for other persons.[11] However, I grant that there are even less sophisticated emotions, which are often reactions to primarily perceptual inputs. What emerges, though, is a sort of spectrum of emotional sophistication with things like primal fear and disgust on the one end and things like intricate forms of moral indignation on the other. If there is to be a differential treatment of sophisticated and unsophisticated emotions, there will be need to be a non-arbitrary cutoff on that spectrum where emotions go from being foundational to non-foundational reasons. This is at least an initial obstacle for such a view. So, while I am not in principle opposed to such an analysis, it would need to be demonstrated what particular feature of an emotion makes it a foundational vs non-foundational epistemic reason.

In summary, because perceptualists treat Link 1 as a merely causal (non-reasoning) link, they are forced to treat evidence of emotional unreliability as an undercutting defeater. As a result, evidence of emotional unreliability (of some emotion type) will *always* defeat emotional justification for beliefs formed on the basis of emotions (of that type). Inferentialists, on the other hand, can treat evidence of emotional unreliability as HOE. Since HOE of unreliability does not directly entail defeat, this opens up more possibilities for interpreting the proper response to evidence of emotional unreliability. In the following section, I explore these possibilities in more depth.

V. Inferentialism and HOE of emotional unreliability

In the previous two sections I argued, first, that not all evidence of emotional unreliability defeats emotional justification and, second, that perceptualists are committed to the claim that all evidence of emotional unreliability defeats emotional justification. I have argued that inferentialists can treat evidence of emotional unreliability as HOE. But, since some think that HOE of unreliability can still defeat justification (albeit by way of a different mechanism than undercutting defeat), I must say more about why I take the inferentialist to be in a better position than the perceptualist. It is this issue that I turn to now.

HOE and defeat are complicated phenomena. I cannot do justice to the many intricacies of the literature on HOE here. Rather than attempting to do so, I will give a big picture overview of the landscape of the HOE debate as it applies to the question of evidence of emotional unreliability. For our

[10] See Ekman (1992).

[11] Indeed, this seems central to the success of Twain's literary choice to use a naïve child narrator.

purposes here, we can sort views of HOE into two categories. The first view says that when you receive HOE that your belief that p was badly formed, you must bracket all of the FOE and reasoning that the belief was based on for the purpose of further reasoning. This is traditionally referred to in the literature as an 'independence principle'. But for our purposes, let's just call this the *bracketing* view. Bracketing is typically defended by proponents of the 'equal weight view' in debates about disagreement and HOE.[12] To deny bracketing is to allow that it is sometimes rational, after receiving HOE that one's belief that p was badly formed, to continue relying on one's FOE and prior reasoning about whether p. Let's call the denial of bracketing *inclusiveness*, since it is inclusive of FOE and reasoning about p.

There are various types of inclusive views. *Steadfasters*, for example, deny that one is rationally required to modify one's belief that p upon acquiring HOE that p was badly formed (Kelly 2005). Aisha, for example, is neither required to reduce confidence that she has enough gas to make it to the destination nor that her belief about having enough gas is justified. Similarly, when Eva receives HOE that her guilt response is unreliable, she remains justified in her belief that she is blameworthy just in case her guilt is, in this case, based on good evidence. This last part is important. No one thinks that *bad* reasons can epistemically justify. This is why it is important in Aisha's case that her FOE (the readings on the gauges, dials, maps, etc.) actually does support her belief that she has enough gas. So, if emotions are based on reasons, as the inferentialist holds, then those reasons must be good ones in order to obtain the emotional justification in the first place. But if that condition is met, then it follows from the steadfast view that Eva does not *lose* her initial emotional justification when she obtains HOE that her guilt is unreliable.

Level-splitting views are also inclusive in this sense. On these views, subjects who receive HOE that their belief that p was badly produced are epistemically permitted to continue believing p but must lower their confidence in the proposition that their belief that p is *justified*.[13] According to level-splitting views, Aisha is not required to reduce confidence that she has enough gas to make it to the destination, but she is required to reduce her confidence that this belief is justified. So too, when Max provides Eva with evidence that her guilt is unreliable, Eva retains her justification for believing that she is blameworthy, but she is required to lower her confidence in the proposition that *she is justified* in believing that she is blameworthy.

On both steadfast and level-splitting views, then, inferentialists can explain why evidence of emotional unreliability does not always defeat emotional justification. It is because evidence of emotional unreliability is HOE, and HOE does not defeat justification (for the first-order conclusion). This allows that

[12] See Christensen (2007, 2010) and Elga (2007).
[13] See Weatherson (2013) and Lasonen-Aarnio (2014).

our three subjects from the counterexamples in Section III, Helena, Hans, and Chuck, all retain their first-order emotional justification. On level-splitting views, their justification for the meta-beliefs that their emotional responses are justified is threatened. But this is unproblematic since what we want to preserve in these cases is the possibility of these subjects rationally treating their emotions as epistemic reasons, not their meta-beliefs about whether or not those emotions are justified.

Both steadfast and level-splitting views allow the subject to, in some sense, ignore the HOE that their first-order belief was badly formed. But some inclusivist views require the subject to be more responsive to their HOE. For example, on Kelly's (2010) 'total evidence view', one needs to be rationally responsive both to FOE and HOE. On this sort of view, what one is rationally required to do in a particular case will depend on the details of the case and what one's total evidence (at all levels) supports. Sometimes, one might be required to reduce confidence in p after learning that one's belief that p was badly formed. Other times, it might be rational not to reduce one's confidence in p. Other times, it might be required that one completely withhold judgment with respect to p.

The total evidence view too is compatible with my take on emotional unreliability. So long as it is sometimes the case that one is permitted to treat one's emotion as a reason upon receiving HOE that the emotion was unreliably formed, then my interpretation of the cases of Helena, Hans, and Chuck is unproblematic. On the total evidence view, even when Chuck receives evidence that his compassion is unreliable, it is plausible that his total evidence, which includes, for example, FOE regarding Jim's humanity-making properties, supports the claim that Jim's interests matter, and so he is rational to continue feeling compassion and to treat that compassion as a reason. It is arguable that in Eva's case, her total evidence may not be so supportive of her emotional conclusion. Is her FOE for feeling guilty really so strong as to withstand the threat of the HOE provided by Max? If not, then Eva is not rationally permitted to continue treating her guilt as a reason for belief. So, on the total evidence view, HOE of emotional unreliability will sometimes defeat emotional justification and sometimes not. And this is compatible with my argument from Section III. Thus, all three inclusive views are compatible with inferentialism and the possibility of emotional justification despite evidence of emotional unreliability.

Now for bracketing views. Recall that bracketing views hold that when one receives HOE that one's belief was badly formed, one must rationally bracket the FOE and reasoning processes that led to the belief. Aisha, for example, would be required to set aside the reasoning that led her from her FOE about the gauges, dials, and maps in her assessment of whether she has enough gas to reach her destination. Bracketing views do indeed seem potentially problematic for the picture of emotional reasoning that I have argued for

here. For they would require Chuck, for example, to set aside all of the FOE on the basis of which he formed his compassion for Jim (e.g. Jim's humanity-making features). And, of course, if one is required to bracket the evidence on which the emotions are based, then one will also be required to bracket the emotion itself.

So, a strict bracketing view that requires one to set aside all of one's FOE does seem to have the result that evidence of emotional unreliability defeats emotional justification. However, it is worth mentioning that bracketing also causes problems for many other classes of belief. This threatens to give rise to an over-generation of unreliability charges. For example, suppose that Margot is bad at math. And suppose that her teacher says to her, 'Margot, your mathematical reasoning processes are unreliable, and so you shouldn't treat any of your mathematical beliefs as reasons'. The upshot would be that in order for Margot to get a justified mathematical belief, she is required to base that belief on the testimony of a more mathematically reliable epistemic agent. Even if Margot is bad at math, to say that once she has evidence of mathematical unreliability, she is *incapable* of forming a (non-testimony-based) justified mathematical belief, and thus that she should treat *none* of her (non-testimony-based) mathematical beliefs as reasons, seems to go too far. (Notice that this point is different from the claim that Margot should revise a *specific* mathematical belief in the face of disagreement with, say, her teacher. I will pick up on this point shortly.)

Proponents of bracketing should aim to rule out the result that Margot is rationally required to suspend judgment on *all* (except testimonially acquired) mathematical beliefs. But if they successfully rule out the mathematical case, it seems plausible that they will also rule out the emotion case. Just as Margot should not be required to bracket all of her mathematical reasoning, neither should we be required to bracket all of our emotional reasoning.

I do not mean to claim that one is *always* epistemically permitted to use one's emotions as reasons for belief. It is especially important to remember that according to inferentialism, one is permitted to treat one's emotions as epistemic reasons only if those emotions are themselves based on *good* evidence. Precisely what it takes to be based on good evidence will depend on one's account of evidence and basing. But we can see how some obviously bad cases will be ruled out by this requirement. For example, if Helena's anger about Joan's being passed over for the promotion is caused by the hatred of men generally and the belief that the selection committee was comprised of men, then her anger would not be based on good reasons, and thus she should not treat that anger as a reason. If Chuck's compassion for Jim is based on an attraction to Jim's physical appearance rather than his humanity-making features, then Chuck's compassion is epistemically unjustified and should not be treated as a reason. It is only when the emotion is based on good reasons that it has any justificatory power to begin with.

In summary, when it comes to HOE and defeat, either (1) inclusive views are correct or (2) bracketing views are correct. If (1), then emotional justification is not undermined by HOE that the emotional response is unreliably produced. Thus, in this case, one is still epistemically permitted to treat one's emotion as a reason on the inferentialist view of emotional reasoning. If (2), then HOE of emotional unreliability would defeat emotional justification, but it would also be the case that HOE of mathematical unreliability would defeat mathematical justification. There would be no special problem, then, for emotional unreliability. Since the mathematical result is so counter-intuitive, it seems to me that bracketing views do not present the inferentialist with a strong reason to worry.

This leaves the inferentialist in a much better position *vis-à-vis* our earlier observation that evidence of emotional unreliability does not always defeat emotional justification. While the perceptualist must treat evidence of emotional unreliability as an undercutting defeater, the inferentialist can treat it as HOE. And so long as one does not adopt a version of the bracketing view of HOE that has strongly counterintuitive consequences (e.g. that Margot cannot trust any of her mathematical beliefs), this allows the inferentialist to explain how emotional justification can go undefeated in the face of evidence of emotional unreliability.

VI. Conclusion

I have argued here for two main claims. The first is that insofar, as one is ready to accept that emotions can provide epistemic reasons for belief, one should also accept that there are cases—such as those of Helena, Hans, and Chuck—in which evidence of emotional unreliability does not defeat emotional justification. Emotional rationalists should aim to preserve this possibility in their theory. The second point is that only emotional inferentialists, and not emotional perceptualists, are in a position to theoretically accommodate this result. This is because treating emotional reasons as epistemically foundational, as perceptualists do, requires treating evidence of emotional unreliability as an undercutting defeater. Treating emotional reasons as non-foundational, by contrast, opens the possibility of treating evidence of emotional unreliability as HOE. And on any inclusivist view of HOE, this in turn allows the inferentialist to explain the possibility of undefeated emotional justification in the face of evidence of emotional unreliability. Taken together, these points provide emotional rationalists with an important argument for preferring inferentialism to perceptualism.[14]

[14]This paper has benefited greatly from comments from Uriah Kriegel on a previous draft, helpful discussions with Kathrin Glüer-Pagin, challenges from anonymous reviewers, and the insightful questions of the participants of Stockholm University's Higher Seminar in Theoretical Philosophy. The work has been supported by Riksbankens Jubileumsfond, grant number M18-0310:1.

References

Arpaly, N. (2003) *Unprincipled Virtue: an Inquiry into Moral Agency*. Oxford University Press.
Bennett, J. (1974) 'The Conscience of Huckleberry Finn', *Philosophy*, 49 123–34.
Brady, M. (2013) *Emotional Insight: the Epistemic Role of Emotional Experience*. New York, NY: Oxford University Press.
Carter, J. A. (2020) 'Epistemic Perceptualism, Skill, and the Regress Problem', *Philosophical Studies*, 177: 1229–54.
Christensen, D. (2007) 'Epistemology of Disagreement: the Good News', *The Philosophical Review*, 116: 187–217.
Christensen, D. (2010) 'Higher Order Evidence', *Philosophy and Phenomenological Research*, 81: 185–215.
Deonna, J. and Teroni, F. (2012) 'From Justified Emotions to Justified Evaluative Judgements', *Dialogue*, 51: 55–77.
Döring, S. (2008) 'Conflict without Contradiction', in G Brun, U Dogluoglu and D Kuenzle (eds), *Epistemology and Emotions*. Ashgate Publishing Company. pp. 83–103.
Ekman, P. (1992) 'An Argument for Basic Emotions', *Cognition and Emotion*, 6 169–200.
Elga, A. (2007) 'Reflection and Disagreement', *Noûs*, 41: 478–502.
Elgin, C. Z. (1996) *Considered Judgment*. Princeton: New Jersey: Princeton University Press.
Goldie, P. (2008) 'Misleading Emotions', in G. Brun, U. Dogluoglu and D. Kuenzle (eds), *Epistemology and Emotions*. Ashgate Publishing Company. pp. 149–65.
Horowitz, S. (2022) 'Higher-order Evidence', *The Stanford Encyclopedia of Philosophy (Fall 2022 Edition)*, E.N. Zalta and U. Nodelman (eds). https://plato.stanford.edu/archives/fall2022/entries/higher-order-evidence/
Kelly, T. (2005) 'The Epistemic Significance of Disagreement', in *Oxford Studies in Epistemology, Volume 1*, T.S. Gendler and J. Hawthorne (eds), Oxford/New York: Oxford University Press, 167–96.
Kelly, T. (2010) 'Peer Disagreement and Higher Order Evidence', in Richard Feldman and Ted A. Warfield (eds), *Disagreement*. Oxford University Press.
Lasonen-Aarnio, M. (2014). 'Higher-order Evidence and the Limits of Defeat', *Philosophy and Phenomenological Research*, 88: 314–45.
Mitchell, J. (2021) *Emotion as Feeling Towards Value: a Theory of Emotional Experience*. Oxford: Oxford University Press.
Nussbaum, M. (2001) *Upheavals of Thought*, Cambridge: Cambridge University Press.
Pollock, J. L. (1987) 'Defeasible Reasoning', *Cognitive Science*, 11: 481–518.
Silva, L. (2022) 'Emotions and Their Reasons', *Inquiry: An Interdisciplinary Journal of Philosophy* 1, pp. 1–24.
Solomon, R. (1977) *The Passions*, New York: Anchor.
Tappolet, C. (2016) *Emotions, Value, and Agency*. Oxford: Oxford University Press UK.
Tappolet, C. (2021) 'Emotions Inside Out: the Content of Emotions', in, and C. Demmerling and D. Schroder (eds), *Concepts in Thought, Action, and Emotion: New Essays*. New York: Routledge.
Twain, M. (1884/1994) *The Adventures of Huckleberry Finn*. New York: Penguin Books.
Weatherson, B. (2013) 'Disagreements, Philosophical and Otherwise', in J. Lackey and D. Christensen (eds), *The Epistemology of Disagreement: New Essays*. Oxford University Press. pp. 54.