CAN MORAL REALISTS DEFLECT DEFEAT DUE TO EVOLUTIONARY EXPLANATIONS OF MORALITY?

BY

MICHAEL KLENK

Abstract: I address Andrew Moon's recent discussion (2016, this journal) of the question whether third-factor accounts are valid responses to debunking arguments against moral realism. Moon argues that third-factor responses are valid under certain conditions but leaves open whether moral realists can use his interpretation of the third-factor response to defuse the evolutionary debunking challenge. I rebut Moon's claim and answer his question. Moon's third-factor reply is valid only if we accept externalism about epistemic defeaters. However, even if we do, I argue, the conditions Moon identifies for a valid third-factor response are not met in the case of moral realism.

1. Introduction

Some moral realists believe that we can have objective moral knowledge: moral properties and facts exist stance-independently, and we can have justified true beliefs about them (e.g. Shafer-Landau, 2003; Enoch, 2011; Wielenberg, 2014). The reliability challenge against objective moral knowledge is intended to show that all of our moral beliefs are unjustified – at least insofar as the moral beliefs are about stance-independent properties and facts of the sort defended by moral realists. The challenge is often based on evolutionary explanations of morality (Street, 2006, 2016; Joyce, 2006).[1] If the challenge succeeds, and if justification is required for knowledge, then objective moral knowledge seems impossible. Moon (2016) understands the reliability challenge to be a probabilistic challenge.

Pacific Philosophical Quarterly 98(S1) (2017) 227–248 DOI: 10.1111/papq.12207

Accordingly, the reliability challenge provides an epistemic defeater for R_m, where R_m stands for the belief our moral beliefs are generally reliable, for anyone who believes that the probability is low that our moral beliefs are reliable, given our evolutionary past and the truth of moral realism (Moon, 2016, pp. 8–9).[2]

Realists seem to have a straightforward answer to this challenge: so-called third-factor accounts assume the truth of some particular, substantive 'Morality Claim' M (Moon, 2016, p. 7), and then use M to demonstrate that our moral beliefs are by and large reliable (e.g. Enoch, 2010; Wielenberg, 2010; Brosnan, 2011; Skarsaune, 2011). For instance, the morality claim favoured by realist David Enoch is that 'survival or reproductive success is at least somewhat good' (Enoch, 2010, p. 430). Faced with a probabilistic version of the reliability challenge, realists might argue that, given that M is true, there is a moderately high probability that R_m is true, too (Moon, 2016, p. 7). Hence, our moral beliefs seem justified after all and therefore the reliability challenge is thwarted.

However, whether realists are entitled to assume that M is true is the key question, if not the 'heart of the debate between realists and the debunker' (Vavova, 2015, p. 111). There are two aspects to this debate. On the one hand, the third-factor response against the reliability challenge appears problematic because the reliability of our moral beliefs is called into question by a defeater – it seems circular or question-begging to rely on a defeated moral belief to fend off the challenge. Many take this to be a reason to dismiss swiftly the realist reply (Street, 2008; Vavova, 2015, p. 111; Fraser, 2014, p. 471). On the other hand, circularity seems, to a certain extent, inherent in explanations of the reliability of any class of beliefs. Arguably, we can explain the reliability of any particular class of beliefs only if we assume the truth of some of the beliefs in question (White, 2010; Berker, 2014). For instance, if you wonder about the reliability of your maths-beliefs, you might work out the 42nd decimal of Pi and check your result against what you believe to be the truth (the answer is 9). Relatedly, some argue that we can trust our cognitive capacities without prior evidence without begging the question (cf. Foley, 2001). Third-factor responses could fall into either category, so they are not quite so easy to dismiss.[3]

The way forward in this debate suggested by Moon (2016) is to identify a response to epistemic defeaters that does not beg the question and to argue that if moral realists can adopt this response, then they are on their way to answering the reliability challenge.

In this article, I address the question that Moon leaves unanswered: is Moon's proposed interpretation of the third-factor response available to moral realists? Can they 'deflect' the possible defeat due to evolutionary explanations of morality? I argue for two points. First, the success of Moon's generic response to defeaters depends on the truth of externalism about epistemic defeaters.[4] However, Moon does not address the relevance of

externalism for his interpretation of the third-factor response and what he writes suggests that he assumes that internalism is true. So, Moon's interpretation of a valid third-factor response is either false, if internalism is true, or misleading, insofar as Moon fails to address the relevance of externalism for his interpretation of the third-factor response.[5] Second, and more directly pertinent to the reliability challenge in the moral case, even if we assume the truth of externalism to make Moon's general strategy work, the particular case of moral realism does not satisfy the conditions for employing Moon's strategy validly.

Although I ultimately reject the applicability of Moon's answer for the moral case, his proposal, and my discussion of it, should be of interest for those who are working on the metaethical debunking debate. It should help with locating the problem with answering the reliability challenge in the moral case and thus clarify the constraints faced by moral realists.

A caveat: I assume for the sake of the present discussion that the reliability challenge (and evolutionary debunking explanations in particular) provides us with a defeater of all our moral beliefs, irrespective of whether the defeater can be dealt with, quite like the defeaters in the examples that Moon discusses. This is a controversial assumption and if realists can show that it is mistaken, then they would not need to rely on MOON'S WAY OUT in the moral case in the first place.[6]

Section 2 introduces Moon's argument. Section 3 contains my criticism of Moon's argument for a valid version of the third-factor response. Section 4 identifies disanalogies between Moon's artificial cases and the metaethical debunking challenge, and in Section 5 I discuss and reject one further twist that might make the analogy between Moon's cases and the reliability challenge work.

2. *Moon's interpretation of the reliability challenge*

Moon's probabilistic formulation of the reliability challenge relies on the notion of an epistemic defeater.[7] Short of an explicit definition, he uses the following case, adopted from Plantinga (2000), to illustrate what an epistemic defeater is:

> XX Pill Case: You learn that a pill, called 'XX', destroys the cognitive reliability of 95% of those who ingest it. You take the pill and come to believe both that I've ingested XX and P(R/I've ingested XX) is low (Moon, 2016, p. 3).

According to Moon, realists have an undermining defeater for R_m – the belief that their moral cognition is generally reliable – if they believe that the probability is low that our moral cognition is reliable if moral realism

is true and moral cognition is an adaptation. Moon assumes that realists do accept that the probability is low that R_m is true, conditional on the claim that moral realism and the evolutionary claim are true (Moon, 2016, p. 9). Adherents of the third-factor response want to use the morality claim to conclude that the probability that R_m is true is at least moderately high. However, Moon notes that in most cases '… it is illicit to use the Morality Claim to prevent the defeat of R_m *when belief in the Morality Claim is a deliverance of the very faculties that are in question*' (Moon, 2016, p. 7; emphasis added). This is the problem in the moral case.

However, Moon argues that it is false to assume that is generally impossible to use beliefs produced by faculty F to avoid defeat of the faculty F. He uses the following case to illustrate this point:

> XX Deflector Case: All is as in the original XX pill case, except that before I took the pill, a scientist I know to be trustworthy had informed me that I am one of the 5% who is immune to the drug. I then take XX while knowing that *I am one of the immune 5% and P(R/I've ingested XX and I am one of the immune 5%) is high* (Moon, 2016, p. 13, italics in original).

In the XX-Deflector case, Moon argues, we can legitimately use a belief (which is obviously a 'deliverance' of our cognitive faculty) to *deflect* a potential defeater of our cognitive faculty. Moon calls this way of using the morality claim a *defeater-deflector* and argues that the 'charge of question-begging' against third-factor replies fails *if* the morality claim is understood as a defeater-deflector (Moon, 2016, pp. 13–14).

Why does the defeater-deflector account succeed according to Moon? He argues that the defeater-deflector account works because the subject in the XX-Deflector case 'never once gain[s] a reason to doubt R'[8]: the subject starts with an undefeated belief in R, gains evidence that he is immune to the pill *prior* to ingesting the pill, takes the pill with the scientist's testimony still vividly in mind, and consequently is aware that he took a drug that he is immune to. Moon contends that this scenario would make it 'very odd' to think that the XX pill could defeat R (Moon, 2016, p. 14). So, Moon argues, *conscious memory* of the scientist's testimony received *before* the ingestion of the pill suffices to deflect the potential defeater gained by the XX pill. Moon emphasises the importance that the subject be 'conscious' of the scientist's testimony after ingesting the pill, which is clear when he considers a turn of events that would cause the subject to lose its defeater-deflector (Moon, 2016, pp. 14, 16):

> […] I lost the belief (due to cognitive decline) that I am one of the immune 5%, but I did continue to believe that I took a drug that 95% of the population is vulnerable to. Then I would clearly get a defeater because I would no longer have the deflector. But so long as I continue to consciously believe that I am one of the immune 5%, it seems that the belief continues to have its deflecting powers (Moon, 2016, p. 14).

The facts about mental states highlighted by Moon, such as whether the subject consciously believes that he is immune to the defeater, are relevant only on internalist views on epistemic defeaters (cf. Grundmann, 2014, p. 157); Moon's emphasis on these introspective facts implies that he is operating on the assumption that an internalist perspective about epistemic defeaters is correct. However, Moon leaves open whether what he calls a 'strong internalist view' or a 'moderate internalist' is correct (Moon, 2016, p. 17).[9] This distinction between both internalist views will be relevant for assessing Moon's account, so let me briefly explain it.

Strong internalism counts only 'present, conscious states' as relevant (hence the emphasis that the subject be aware of the scientist's testimony) while moderate internalists takes 'unconscious states' to be 'justificationally relevant', too (Moon, 2016, p. 17). A moderate internalist will say that the subject has a defeater-deflector even if the belief is not conscious, but nevertheless accessible (Moon, 2016, p. 17).[10] Figure 1 depicts the sequence of events that are crucial in Moon's XX-Deflector case[11]:

The subject S does not have a defeater prior to t_3. The decisive point is t_4. At t_4, S has already learnt about the devastating effects of the pill and ingested the pill. According to Moon, S possesses a defeater-deflector at t_4. Moon suggests that, depending on whether strong or moderate internalism about defeaters is true, S continues to possess a defeater-deflector at t_5 but not at t_6 (strong internalism) or both at t_5 and t_6 (moderate internalism). Since Moon leaves both interpretations of internalism open, we can summarise Moon's principle about the sufficient conditions for responding to epistemic defeaters as follows:

MOON'S WAY OUT: A subject S's belief f processed by faculty F can deflect the potential defeater PD of faculty F if f is formed prior to the reception of PD and, if strong internalism is true, S is conscious of f, or, if moderate internalism is true, S can access f.

Moon does not claim that the defeater-deflector interpretation expressed by MOON'S WAY OUT is the *only* escape route from the reliability challenge

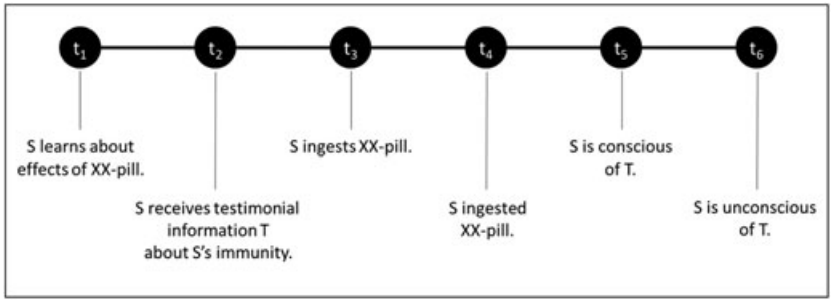

Figure 1. XX-Deflector case: sequence of events.

for realists. However, since he argues that two popular existing alternative interpretations of the third-factor account fail, he does regard the defeater-deflector interpretation as 'the best way to understand the third-factor response' (Moon, 2016, p. 11).[12]

But can moral realists use a claim in accordance with MOON'S WAY OUT in the moral case? Moon records a conditional answer: *if* realists are in a situation similar to the XX-Deflector case, then they can use his interpretation of the third-factor reply. However, Moon is unsure about the antecedent of the conditional (Moon, 2016, p. 14). Call this MOON'S QUESTION:

MOON'S QUESTION: Can moral realists use a defeater-deflector as in the XX-Deflector case?

Moon records that he does 'not know' the answer, but he surmises that realists might be unable to use the defeater-deflector (Moon, 2016, p. 14). He also offers a contrasting case that, if it resembles the realist's situation, bodes trouble for a successful third-factor response because it does not contain any beliefs that 'have the power to deflect' defeaters (Moon, 2016, p. 12):

YY Colour Vision Deflector Case: [A pill, called 'YY' renders unreliable the colour vision of 95% of those who ingest it.] You do not know about YY, and you find yourself in [a room with objects that have no standard colour. For example, there are no bananas or blue jays, but there are plastic bowls, walls, and a chameleon]. You have already formed beliefs that the wall is red, that a bowl is white, and so on. Then a friend who you know to be a very reliable testifier tells you about YY and that YY was mixed into the dinner you had enjoyed earlier that evening. You come to believe that 'I ate YY' and that 'P(RC/I ate YY) is low.'

I will come back to the YY-Deflector case in Section 4.3; the important features of the case are that, in contrast to the XX-Deflector case, it does not as such contain a defeater-deflector and the defeater affects only one particular belief-forming faculty, as opposed to all belief-forming processes.

This completes Moon's picture as it is relevant to my discussion. The two crucial points are MOON'S WAY OUT (realists can respond to epistemic defeaters if they have a defeater-deflector), and MOON'S QUESTION (do moral realists have a defeater-deflector?). If MOON'S WAY OUT were true, it would provide a blueprint for responding to the most menacing epistemic defeaters. In what follows, I will challenge MOON'S WAY OUT and provide a negative answer to MOON'S QUESTION.

3. *MOON'S WAY OUT is false, or Moon's presentation of it is misleading*

In this section, I argue that the truth of MOON'S WAY OUT depends on the truth of externalism about epistemic defeaters.[13] However, as we have seen, Moon

emphasises that a subject's introspectively available information is important for assessing whether the subject has a defeater, which implicitly suggests that he operates on the assumption that internalism is correct. Hence, if internalism is true, then MOON'S WAY OUT is false. If internalism is false, then MOON'S WAY OUT might work, but then Moon's emphasis on internalist aspects of his argument for MOON'S WAY OUT is misleading.[14]

The problem with MOON'S WAY OUT is that XX-pill debilitates the reliability of the subject's cognitive faculties entirely. Cognitive faculties include memory. So, how could the subject be sure, at t_4, that his memory about his immunity to the pill is itself not just an effect of the pill?

Suppose we are dealing only with users of the pill who are *immune*. Let IMMUNE be a subject's belief that they are immune to the pill. Moon allows them to carry on as usual if they only keep in mind that they are immune to the pill, a concession that seems to rely on an odd picture of human psychology. For why is it important that users of the pill keep in mind[15] that they are immune to the belief pill?[16]

Moon does not explain why the mental states of the subjects matter when thinking about the defeater-deflector, but it seems that the *content* of IMMUNE alone cannot be relevant. Imagine a subject that is immune to the pill who, in contrast to the case outlined above, hears about his immunity only after ingesting the pill. From the subject's perspective, the belief in the reliability of his cognitive faculties would already be defeated, so he should doubt subsequent testimony, or memory about testimony, about his own immunity, too. Hence, the content of IMMUNE alone does not suffice to deflect defeat.

Therefore, the important thing about IMMUNE is not its content per se, but having been acquired *before* the ingestion of the pill. We can call this property of the belief TIME-STAMPED: any belief that is TIME-STAMPED was acquired *before* the ingestion of the pill.

Now, if subjects in the XX-Deflector case could identify all TIME-STAMPED beliefs, then they could use the TIME-STAMPED belief IMMUNE to deflect the potential defeater of the XX-pill, as suggested by Moon. Users of the pill would have to be capable of running through their beliefs, putting the 'good', i.e. TIME-STAMPED ones into the pot of reliable beliefs to act on, and discarding the potentially corrupted ones, just like the pigeons in the fairytale 'Cinderella' can pick good from bad lentils.[17] In that case, we could assume that immune users of the pill can introspectively and reliably detect the TIME-STAMPED beliefs.

However, we are no pigeons and, more importantly, our beliefs are no lentils. Lentils can be looked at to identify markers of quality carried by them. The same is not true of beliefs. Introspectively, the property TIME-STAMPED for the belief IMMUNE can only be identified by memorising a meta-belief about the belief IMMUNE. Undoubtedly we do form relevantly similar meta-beliefs in lots of cases.[18] But the important question is whether

such meta-beliefs are also introspectively reliable: can immune users rely on the meta-belief that <My belief IMMUNE is TIME-STAMPED >? It is far from obvious that they can. First, coming to believe in IMMUNE would have to be a sufficiently salient event to stimulate the formation of a related meta-belief about TIME-STAMPED. This might seem likely, given the potentially disastrous effects of the XX-pill, but the salience also depends on the subject's estimate of the likelihood that he will later actually ingest the XX pill, about which we lack information in the XX-Deflector case. Second, once the subject the pill, we should expect him to believe that IMMUNE is TIME-STAMPED because that would allow him to continue believing the functioning of his cognition. It is thus likely that his meta-beliefs would be liable to confirmation bias (cf. Nickerson, 1998). Third, it is unclear whether subjects could continuously keep IMMUNE and the meta-belief about TIME-STAMPED conscious, as required by the 'strong internalist' scheme that Moon considers as a candidate position about the epistemology of epistemic defeaters. Rather, it seems likely that subjects will soon let both beliefs slip from their consciousness and thus, on strong internalism, their possession of a defeater-deflector seems at best short-lived. Fourth, on a 'moderate internalist' interpretation, we could accept that subjects are temporarily unconscious about TIME-STAMPED, but the retrieval and storage of information in memory makes meta-beliefs particularly liable to corruption (cf. Schacter *et al*., 2011); hence their merely accessible beliefs about TIME STAMPED tend to become less reliable over continued recollection.

These are good reasons to suspect that users of the XX pill should not rely on their meta-beliefs about TIME STAMPED. But this intuition might not be shared, and I suspect that one reason for resisting this intuition is because in judging the case we note that users are, in fact, immune to the pill. In light of this fact there might seem to be, from *our perspective*, no reason for the users of the pill to give up their belief in the reliability of their cognitive faculties.

However, Moon's discussion only leaves us with a choice between strong or moderate internalism. So, we have to imagine the situation as it looks from the perspective of a user of the pill. From the internalist perspective, we cannot incorporate the fact that some users are justified to rely on beliefs about their immunity while others are not. Suppose Anton is immune to the pill while Bert is not. Anton believes that IMMUNE is TIME-STAMPED. But Bert does too. Bert, contrary to fact, vividly remembers someone telling him about his immunity to the pill and he recalls that he got told just one week before he ingested the pill. The belief that <my belief IMMUNE is TIME STAMPED > will seem perfectly veridical for both Anton and Bert. Judged from their introspective capacities alone, we have no reason to suppose that their meta-beliefs are phenomenologically different and – importantly – no reason to suppose that Anton is in a better position to rely on his meta-belief than Bert is. Bert might very well believe that he is immune, he might very

well remember that he spoke to a scientist who told him so. Although this is a false belief, his subjective experience will likely be very similar to that of Anton who is, in fact, immune to the pill.

Therefore, users of the XX pill should not rely on beliefs that seem reliable *to them*. With a drug as pernicious as the XX pill, decisions about which beliefs to doubt and which to trust should be left to those who did not take the drug. This suggests how we could resolve the stalemate between Anton and Bert: by pointing out that Anton is, in fact, immune to the pill, while Bert is not. But this is certainly impossible according to strong internalism, which requires us to evaluate how mental properties of the subject influence our judgement about the presence of a legitimate defeater-deflector.

Does adopting 'moderate internalism' rescue Moon's analysis? Recall that moderate internalists hold that unconscious mental states are justificationally relevant, too. In the figure above, the subject is conscious of his immunity to the pill right up until t_5 but unconscious of it at t_6. Still, according to moderate internalism the subject would have a defeater-deflector at t_6 if the information is still present and accessible in the subject's memory. In that case, Moon writes, moderate internalism implies that the 'unconscious belief still has deflecting powers, despite its being unconscious,' since it is still *accessible* (Moon, 2016, p. 17). In a related paper, Moon suggests that *accessibility*, and not consciousness of, is indeed the hallmark of an *internal* property, since he defines an internal property as internal if and only if it is 'introspectively accessible to [a subject]' (Moon, 2012, p. 347). However, the problem seems to be entirely independent of the debate between strong and moderate internalists. They might quibble over what happens at t_6 in the graph above, but the real concern is with whatever happens from t_4 onwards.

The only way I see to rescue MOON'S WAY OUT is by adopting an externalist perspective on defeaters. The externalist view entails that defeaters are true propositions which need not be known to the agent in question. In that case, we can square MOON'S WAY OUT with the intuition that the subject's memory is called into question by the XX pill. The externalist perspective has it that, as a matter of fact, the subject is immune to the pill throughout; hence when we who are not affected by the pill judge the case, we can rely on this information to conclude that the information about the subject's immunity deflects the potentially defeating information about the effects of the pill.[19]

Therefore, MOON'S WAY OUT is either false or incoherent. It is false if we accept internalism about defeaters as true because it seems that internal properties do not afford an introspective difference between subjects that are immune and subjects that are not immune in the XX-Deflector case. We can make sense of MOON'S WAY OUT by adopting externalism about defeaters, but then Moon's repeated emphasis on aspects of the XX-Deflector case that are relevant only from an internalist perspective seems incoherent.

We can already draw an important conclusion for moral realists. Based on the assumption that MOON'S WAY OUT is indeed the best hope for moral realists to respond to the reliability challenge, it seems that the legitimacy of the third-factor depends on externalism being correct.

It gets worse for realists: in what follows I address MOON'S QUESTION and assess whether moral realists can use MOON'S WAY OUT. To do so, I assess whether moral realists are in a situation similar to the XX-Deflector case. In my view, no matter how we understand the crucial feature that makes the XX-Deflector case contain a defeater-deflector that feature is not found in the case of the reliability challenge against moral realism. It appears that not even a cognition-destroying drug as in the generic XX-deflector case is a proper match and analogy for the dire epistemic situation of moral realism.

4. *Answering Moon's question – The XX-deflector case is not analogous to the reliability challenge*

Suppose we agree with Moon that the subject's meta-belief about the scientist's testimony is reliable (either because we believe that externalism is correct or because we accept Moon's argument about the relevance of some information being *conscious* or *accessible*).

Even so, to take MOON'S WAY OUT in the moral case, we would have to accept another extremely controversial assumption: namely, that the scientist's testimony is reliable in the first place. In the context of the XX-Deflector case, the assumption might be benign. But it gets controversial when we regard the XX-Deflector case as an analogy to the reliability challenge in metaethics, where the big question is precisely *whether* there is a reliable source of information at all.

Moon seems to suggest three reasons to assume that there is a relevant source of reliable information available in the XX-Deflector case. But none applies to the metaethical reliability challenge.

4.1. TOKEN AND TYPE

First, Moon suggests that we can distinguish between defeated and reliable *tokens* of belief-forming faculty *types*. That is, we worry about the cognitive faculties of those who ingest the XX pill but not about the cognitive faculty of the scientist who testifies about the subject's immunity to the pill. However, the scientist uses his own cognitive capacities to form, process, and utter this judgement. In most cases, we have no reason to suspect that the scientist's belief-forming faculty is unreliable. This is because the scientist's *token* of the faculty-*type* 'cognition' is reliable (at least that is what we assume), while the tokens of the cognitive-faculty of all XX-pill users are in jeopardy. To illustrate, suppose we both taste sugar and agree about its

sweetness. Then I give you what is commonly known as the 'miracle berry' (*Synsepalum dulcificum*) which causes sour food to taste sweet, while sweet food still also tastes sweet. We then taste a white substance that could be either fine-grained sugar or lemon juice powder. In this case, where we both know that your taste-faculty is disturbed, I can still rely on my taste-judgement although it is based on a token of the faculty-type that is disturbed in your case. There is no problem because the potentially disturbing effect is not *global*. It affects some, but not all tokens of a faculty-type.

The reliability challenge, however, is very different. We are dealing with a defeater of a certain *type* of faculty, namely our moral cognition. There is no reason to suspect that only selected individuals, or tokens of certain types of belief-forming faculties, are unreliable. Instead, we must assume that all of our faculties are unreliable. In other words, the reliability challenge does not apply to tokens of faculties individuated on a personal basis, but rather to a specific type of faculty and the respective objects of the belief in general. In that case, the XX-Deflector case is not analogous to the moral reliability challenge; one condition that might allow for the existence of a defeater-deflector in the XX-Deflector case is not given in the moral reliability challenge.

Realists might respond that there are moral experts whose moral cognition is unperturbed by potential defeaters. However, while there might or might not be moral experts, we should then ask *why* those moral experts have reliable moral cognition. Unless we accept the reliability of moral experts as beyond doubt, as rock-bottom in our attempts to justify moral beliefs about objective moral facts, we cannot merely take their supposed reliability for granted.[20]

Therefore, individuating reliable and unreliable faculties on a token basis does not succeed in the case of the moral reliability challenge. We should not assume that certain individuals will have reliable information, as opposed to others, and so we should not expect this feature of the XX-Deflector case to be analogous to the moral reliability challenge.

4.2. BEFORE AND AFTER

Second, Moon suggests that the deflecting information is unimpaired by the defeater because it *predates* the potentially defeating information. This suggests a distinction between information received (and memorised) *before* the reception of a potential defeater and *after*. Whether or not this criterion is ultimately convincing in determining the presence of a defeater-deflector in the XX-Deflector case, it is not met in the case of the reliability challenge.

In the earlier graph, we can see that the subject hears about his immunity *before* ingesting the pill with the known consequences. We might say that the potentially deflecting information is in some sense stored away safely within

the subject's memory before the potentially defeating information is doing any damage. Again, this temporal distinction makes sense in a large number of cases that involve impaired information processing and decision making. For instance, when I sign a contract then it matters whether I got awfully drunk shortly before I signed or soon after. The latter should not worry us at all, legally speaking.

However, if MOON'S WAY OUT is supposed to work in the moral case, then the temporal characterisation of defeater-deflectors is unhelpful. MOON'S WAY OUT suggests that we should be on the lookout for defeater-deflectors that predate potential defeaters. But there is no *prima facie* reason to suppose that there are some domains in which prior-to-defeat deflection is possible and others in which it is not. While it is entirely conceivable that some situations involve appropriately timed deflectors (for example, those in which we talked to reliable informants earlier in the day) and situations in which we do not, this distinction breaks down in the case of potential defeaters that have no determinate point of reception (or ingestion). If there is a problem with our moral cognition of objective moral facts, then we did not acquire this issue at a particular moment in time. Many construe the reliability challenge as an evolutionary challenge, and it seems clear that we are not affected by evolutionary forces like we are by the ingestion of some drug. Moreover, there are good reasons to conceive of the reliability challenge as an *a priori* challenge, and in this case, it is even clearer that it does not arise at a particular moment in time (cf. Klenk, 2017).

Thus, there is no reason to suppose that we can deflect the moral reliability challenge by relying on information that predates the reliability challenge, because the reliability challenge, in contrast to the XX-Pill, does not create a problem at some determinate point in time.

4.3. LOCAL AND GLOBAL

Third, Moon suggests that there is an epistemically relevant difference between the XX-Deflector case and the YY-Deflector case, which can be interpreted as a difference between *global* defeat of all cognitive faculties and *local* defeat of only some cognitive faculties.[21]

The YY-pill, remember, destroys the reliability of the colour perception of most people who ingest it. According to Moon, the YY-Deflector case does not contain a defeater-deflector, so third-factor replies fail if realists are in a situation like the YY-Deflector case. The straightforward difference to the XX-Deflector case is, as Moon suggests that the XX-Deflector case contains a defeater-deflector. As I argued above, this is mistaken, but even if we accept it for the sake of argument, this is not very informative as a contrast to the YY-Deflector case. We have to ask *why* there is a defeater-deflector in one case, but not in the other.

The second distinguishing criterion is that the XX-Deflector case contains a source of relevant, reliable information that provides the subject with (arguably) deflecting information before receiving a defeater; the YY-Deflector case lacks both features.

Suppose we heed Moon's advice and take both cases as blueprints to check whether moral realists are in a situation similar to the XX-Deflector case (in which they could hope, in line with MOON'S WAY OUT, to fashion a valid third-factor response) or rather in a situation like the YY-Deflector case (Moon 2016, p. 14). If the YY-Deflector case were that weak, that exercise seems futile because the question would just be the familiar one of whether the realist is in a situation that affords a defeater-deflector (like the XX-Deflector case) or in a situation where there is just plain defeat. That is the question that many scholars are currently asking, but since Moon suggests a way forward in the debate, we should construe the intended contrast as a bit more nuanced so that we understand *why* there is a defeater-deflector in one but not in the other case.

Fortunately, as Moon recognises himself, the YY-Deflector case can be amended by supposing that relevant and appropriately timed reliable information about one's immunity to the YY-pill exists (Moon, 2016, p. 13). If the YY-Deflector case is amended accordingly, we find that the affected subjects get testimonial evidence that they are immune to the effects of the YY pill and that their colour vision will not be impaired by it. Moon writes that his YY-Deflector case is similar to cases discussed by Street (Street, 2008, p. 216) and Locke (Locke, 2014, p. 231). While the specifics of these cases do not matter here, their structures are instructive. In both Street's and Locke's case, someone has beliefs about a particular subject matter and learns that these beliefs are caused by an unreliable source; hence it seems that the subject receives a defeater for these particular beliefs. Street's and Locke's cases further entail the stipulation that there is no source of relevant information available to influence the erroneous beliefs that stem from the corrupted source.

This, however, is a crucial difference to the XX-Deflector case, which entails that *all* beliefs might be defeated, while the cases of Street and Locke entail that only beliefs formed through a *particular* process, source, or faculty about a *particular* topic, object or event are defeated. To illustrate, imagine that you got all your beliefs about witches and sorcerers from the *Malleus Maleficarum*, a superstitious book written by a Dominican monk in 1486 about how to identify and deal with witches and wizards. Some believe that it influenced witch-hunting in Europe and later the US; be that as it may, we can be quite confident that it contains no truth whatsoever about witches and sorcerers. Similarly to Moon's original YY-Deflector case, and the related cases found in the literature, we have a case where all beliefs about particular objects or events are formed based on a single source: in this case the *Malleus*.

But we can amend all cases by imagining that there is another source of *relevant* information present. It is relevant in that it provides reliable information about the objects or events that your *corrupted* beliefs are about. For instance, you might consult a historian, or indeed just about any sane person living in the 21st century to update your beliefs about witches and sorcerers so that they are reliable.

The XX-Deflector case involves a potential defeater that affects, by stipulation, all of your belief-forming capacities while the cases of Street, Locke, and my *Malleus* case involve potential defeaters that affect only a particular source of your beliefs.[22] We can call the former a *global* defeater, the latter a *local* defeater. The local defeater in Moon's YY-Deflector case defeats one's colour vision, but none one's remaining cognitive skills. Hence, someone who ingests the pill can make use of his non-defeated faculties to marshal a defeater-deflector: he can use his memory of his friend's testimony about his immunity to the YY pill as a defeater-deflector. Once he ingests the pill, he knows that his colour perception might not function properly anymore, but his memory of the scientist's testimony is clearly unimpaired: he knows that his colour vision is fine.

This turns Moon's tentative appraisal on its head: contrary to what he suggests, the XX-Deflector case appears to be the difficult case for the realist, and the (amended) YY-Deflector case seems like the realist-friendly case. In this case, Moon's discussion indeed points to a crucial difference between domains where we can use defeater-deflectors and domains where we cannot: if there is *global* defeat, as in the XX-Deflector case, then we cannot, but if the defeater is only *local*, then defeater-deflectors may be within reach of realists, even if the temporal and agent-based distinctions discussed in Sections 4.1 and 4.2 fail.[23] The relevance of the distinction between local and global defeaters depends on three assumptions. First, that we can validly individuate belief-forming processes (such as a belief-forming method for colour-beliefs, moral beliefs, etc.). Second, that some defeaters affect only particular belief-forming methods or sources of beliefs, such as beliefs formed by cognitive processes or beliefs formed based on the contents of a particular book. Third, that beliefs produced by one faculty or based on a particular source may provide information about the reliability of beliefs formed by another belief-forming faculty or source. To illustrate: taking the pill that destroys your colour vision defeats your perceptual beliefs about colour, but not your memory that you are immune.

Recall that we assumed that evolutionary considerations would provide a defeater for all beliefs produced by moral cognition. Hence, realists could use the defeater-deflector for their third-factor response only if they can tap into a source of information that is both distinct from (i.e. not a product of) the deliverances of moral cognition and yet indicative of the reliability of moral cognition.[24]

5. *Non-moral beliefs about the adaptiveness of moral beliefs cannot vindicate moral beliefs*

In the remainder of this article, I consider whether evolutionary considerations, which are distinct from the deliverances of moral cognition, might provide information that could vindicate the reliability of moral cognition, although I ultimately reject this proposal.

The thought is as follows: even if all moral beliefs would potentially be defeated by evolutionary explanations of our moral beliefs, we might be able to defend the reliability of our moral beliefs in reference to *non-moral* beliefs that we have no reason to regard as unreliable. The blueprint for such an approach can be found in evolutionary vindications of the reliability of our beliefs in other domains of inquiry, such as epistemic beliefs (de Cruz *et al.*, 2011), beliefs about logic (Schechter, 2013), or our perceptual beliefs (Boudry and Vlerick, 2014). These evolutionary vindications make, roughly, the following points: evolutionary considerations show that the beliefs under scrutiny had to be *true* to be adaptive. On a representational model, this means that the relevant beliefs had to correctly represent the world. If our ancestors held true beliefs for these evolutionary reasons, then we currently hold sufficiently many beliefs, or derivatives of beliefs, whose truth is plausibly seen as adaptive. So, our belief-forming methods in the relevant domain are plausibly regarded as reliable.

Now, I do not wish to discuss the merits of these accounts here. What I want to point out is that if we regard considerations about the adaptiveness and truth-conditions of moral beliefs as non-moral considerations, then realists might gain independent evidence, similar to the scientist's testimony in the XX-Deflector case, about the reliability of our moral beliefs. Realists would be on their way to out of the reliability challenge.

However, such an evolutionary vindication is problematic and unlikely to succeed in the moral case, for at least two reasons. First, moral realists typically do not aim for an empirical vindication of their accounts, nor do they accept the relevance of an empirical vindication of their claims (e.g. Enoch, 2011). On the contrary, the moral realists with whom I am concerned in this article argue that moral properties are of an altogether different sort than natural properties. Hence, while realists might hold that our beliefs *about* moral properties might have played an evolutionary role, they would not require that moral properties themselves played any relevant role in our evolutionary history. It is a strange development (and one which somewhat betrays their theoretical commitments) that moral realists should nonetheless try to come up with reasons for thinking that moral properties played a causal role in human evolutionary history.

Second, and more important, realists would have to show that evolutionary considerations vindicate the evolutionary relevance of the *truth* of our

moral beliefs. But this seems unlikely to be successful because the truth of our moral beliefs was unlikely to have played an evolutionary role. To see this, consider that claims to the effect that *true beliefs produced by faculty F were adaptive* imply three points. First, that beliefs formed by faculty F generally represented the world to be in a certain way; second, that the probability of certain actions increased due to these beliefs; and, third, that the combination of representation of the world and the ensuing reaction was such that it conferred an evolutionary advantage to our ancestors. Thus, actions, not beliefs, are the primary factor in considerations about adaptiveness. The general point is that the content of beliefs matter in evolutionary explanations insofar as we can show that certain world-to-belief relations are more likely to lead to adaptive actions than others. The implications for moral realists is that beliefs matter in evolutionary explanations insofar as they might be able to show that, more often than not, *true* representations of the world were more likely to lead to adaptive actions.

However, actions purportedly guided by moral beliefs, such as speaking the truth or taking care of one's offspring, appear to be adaptive quite irrespective of whether their related moral beliefs correctly represent the world (cf. Gibbard, 2003, ch. 13; Street, 2006; Joyce, 2006).[25] In other words, there is no need to claim that our moral beliefs were true, i.e. that they correctly represented the moral facts, to explain their potential evolutionary adaptiveness. While this does not imply that our moral beliefs are *false,* neither does it imply that evolutionary explanations show that our moral beliefs are *true*.

What's worse for realists, as long as it is possible to account for the adaptiveness of a moral belief without referring to its truth, the their intended evolutionary vindication of our moral beliefs cannot succeed. Evolutionary considerations might show that realism is *compatible* with evolutionary explanations of our moral beliefs, but they do not vindicate it. Hence, evolutionary considerations about the adaptiveness of our substantive moral beliefs do not give us reasons to believe that our moral beliefs are reliable.

The problem persists for evolutionary considerations about the adaptiveness of our meta-ethical beliefs, which is another way for realists to approach the task of using non-moral beliefs to vindicate their moral beliefs. The thought might be that our beliefs about, say, the nature of moral properties, really do make a behavioural difference only if they are true. The first step in the argument resembles familiar evolutionary explanations of beliefs in absolute objective standards. Such beliefs arguably turned humans into better co-operators (cf. Tomasello, 2016). Imagine two people; one believes that <cheating is wrong because it gets you into trouble if you get caught>, while the other believes that <cheating is wrong because that is a moral rule>. Plausibly enough, we can imagine that the beliefs of both agents affect their individual behaviour in different ways. Perhaps the one worried about getting caught will cheat if there is no way for getting caught, while the other, holding a characteristically 'objectivist' conception of morality,

might avoid cheating even if there is no chance of getting caught. These considerations suggest that having objectivist metaethical beliefs might matter in an evolutionary sense. Albert Camus' titular character in *The Stranger* is a case in point: the Stranger thinks value is only ever contingent and, surely enough, partly because of this view he ends up with his head under the guillotine, and no time to procreate. Hence, there seems to be a good case for arguing that different metaethical beliefs might have produced actions of differing adaptiveness and, in particular, that characteristically objectivist metaethical beliefs lead to more adaptive behaviour.

However, apart from the fact that the empirical case to back up this intuition is harder to make (for instance, there is little evidence that beliefs in intrinsic values promote cooperative action; reputational effects are more reliable predictors of proxies for moral behaviour, cf. Haley and Fessler, 2005), those metaethical beliefs would be adaptive irrespective of their truth. It *may* turn out that moral objectivists act in ways that increase their relative fitness. But, as in the case of substantive moral beliefs, this is compatible with the claim that these beliefs are all false. Indeed, that is the core point of some debunkers (cf. Ruse, 1998): having objectivist moral beliefs would be adaptive even if those beliefs were false (understood on a correspondence model of truth, cf. Joyce, 2016, pp. 154ff). Therefore, it seems that using non-moral beliefs to deflect defeaters of our moral beliefs does not provide a way out of the reliability challenge for moral realists either. In that case, the answer to MOON'S QUESTION is that moral realists are not in a situation relevantly similar to the XX-deflector case; hence they cannot take MOON'S WAY OUT.

6. *Conclusion*

I carried on with Moon's project to make progress in the debate about the viability of third-factor responses against the metaethical reliability challenge. Moon suggested, through the principle that I called MOON'S WAY OUT, that defeater-deflectors are a valid response to the reliability challenge. He left open whether moral realists can make use of third-factor replies so understood.

I argued that we should believe that there is a defeater-deflector in the case discussed by Moon only if externalism is true. Next, I argued that the conditions for a defeater-deflector are not given in the situation of the moral realist, even if we assume that externalism is true. MOON'S WAY OUT does not work, if it works at all, for moral realists. Hence, moral realists cannot deflect defeat from evolutionary explanations of morality.

For my investigation of Moon's interpretation of the third-factor response, I assumed two points: first, that the defeating power of the reliability challenge is similar to the defeating power of Moon's XX-Pills.

Second, that Moon's arguments against two alternative interpretations of the third-factor account are sound. Realists might challenge both assumptions, either by coming up with alternative takes on the third-factor response, or by showing that the reliability challenge does not give us an epistemic problem on par with cognition-destroying pills. So all is not lost for moral realists.

However, if my assumptions hold true, then the considerations offered in this article suggest that the third-factor response in the moral case is unlikely to succeed. What might have seemed like a way out of the reliability challenge turns out to be a bad moon rising for moral realism.[26]

Department of Philosophy and Religious Studies
Utrecht University

NOTES

[1] In this article, I adopt Moon's understanding of moral realism, according to which moral realism entails three claims: there are stance-independent moral properties and facts; it is possible to have knowledge about these properties and facts; moral properties are irreducible to non-natural properties. This conception of moral realism excludes moral naturalists, such as Brink (1989), who do not think that moral properties are irreducible, and realists who do not think that moral properties are stance-independent, such as Railton (1986). This restriction makes sense because Moon's probabilistic interpretation of the reliability challenge entails the claim that realists have reason to believe that the probability is low that we have reliable moral beliefs, given their conception of morality and evolutionary explanations of morality. It is less clear, and worth a separate discussion, whether non-robust realists have equally strong reasons to believe that that probability is low (cf. Moon, 2016, pp. 8–9).

[2] Moon's overt project in Moon, 2016, is the comparison between the evolutionary argument against naturalism and the evolutionary argument against moral realism. I do not discuss this part of his paper in this article. Instead, I focus on the implications that Moon so helpfully draws from his comparison.

[3] Provided that proponents of the reliability challenge want to be sceptics about morality only and avoid scepticism about logic, mathematics, science, and the external world, they have two options. Either they avoid the problem of reliability in certain domains, for instance by adopting constructivism and/or a deflationary theory of truth in these domains, or they show why circularity in explaining reliability is a problem particularly in the moral case, but not in any of the other cases. Lest we are interested in trite burden shifting arguments (i.e. do moral realists have to show that the moral case is realist-friendly or do moral sceptics have to show that it is realist-adverse?) it is worthwhile to identify 'realists-friendly domains' in the sense that non-question begging defences of our beliefs in that domain, realistically construed, are possible.

[4] Externalism about epistemic defeaters implies that the defeating information need not be accessible (or conscious) to the subject, whereas internalism requires defeaters to be somehow accessible to the subject.

[5] Moon does not provide an explicit definition of epistemic defeaters in Moon (2016) at all. However, as I show below, Moon's discussion *implies* that an internalist conception of defeaters is correct, which is all I need for the claim that it is misleading to suggest a view that works only if externalism is true, while presenting it in accordance with internalist principles.

[6] Some claim that evolutionary considerations do not provide a defeater in the first place (e.g. Mogensen, 2016; Hanson, 2016). However, it stands to reason that the defeating power

of the reliability challenge does not depend on evolutionary causal considerations per se (cf. Klenk, 2017), in which case the defeating power of the reliability challenge seems relevantly similar to the defeating power of the cognition-disrupting pills discussed by Moon. For reasons of space, I cannot engage further in this debate in this article. Thanks to an anonymous referee for prompting me to emphasise this point.

[7] Defeaters are often distinguished into rebutting and undermining defeaters (cf. Pollock, 1970). Moon is concerned with the latter: they affect the justification of a belief, but do not show that the belief is false.

[8] R stands for the belief that our cognitive faculties are generally reliable (Moon, 2016, p. 1).

[9] The fact that Moon draws a contrast between two forms of internalism, rather than between internalism and externalism, lends further support to the impression that he operates under the assumption that internalism is correct. As I stated in the introduction, however, Moon does not explicitly endorse either position, and my claim is only that he does not seem to acknowledge the relevance of externalism for his interpretation of the third-factor response.

[10] Like your belief that you live in house number so-and-so is 'conscious' now but was unconscious but nonetheless accessible before I primed you to think of it.

[11] The depicted distance between events is not representative, nor does it matter for present purpose. The order of events, however, matters. t_4 can be arbitrarily close to t_3 but it must occur after t_3.

[12] I do not assess the claim that MOON'S WAY OUT is the *best* response to epistemic defeaters in this article. However, given that several authors reject the third-factor response as hopelessly question-begging (e.g. Street, 2008; Fraser, 2014; Vavova, 2015, or Joyce, 2016), and Moon's critical analysis of alternative interpretations (Moon, 2016, pp. 10–13), it seems safe to say that a failure of MOON'S WAY OUT in the moral case would indeed be bad news for realists. Conversely, if MOON'S WAY OUT *were* true and applicable in the moral case, then realists should be able to provide a valid third-factor response that avoids the common criticisms of circularity or begging the question.

[13] I am concerned with MOON'S WAY OUT only and my argument touches only on the implications of either externalism or internalism for MOON'S WAY OUT, without claiming to suggest anything about the truth of either position.

[14] I discuss the comparability of the XX-Deflector case and the reliability challenge for moral realists in Section 4. The present section is nonetheless relevant for the metaethical debate: if MOON'S CLAIM is false even in the generic XX-Deflector case, then realists would be hard pressed to find another valid way of responding to the reliability challenge, provided that Moon is right that the defeater-deflector interpretation is the best interpretation of the third-factor response.

[15] That is, at least have the belief *accessible*.

[16] Distinguish four states of beliefs: 'conscious', 'accessible', 'inaccessible', 'absent'. Following Moon's choice of terminology, a belief is conscious at the time a subject is considering it. It is accessible if the subject could bring it to consciousness through wilful effort. It is inaccessible if there are traces of information present that cannot be wilfully accessed by the subject. Other calls this 'implicit' memory. It is 'absent' if there is no information present. Moon does not discuss the latter two cases.

[17] The version of Cinderella recorded by the Brothers Grimm has Cinderella, who lives with her evil stepmother, faced with the task of cleaning lentils while her two evil stepsisters enjoy the king's festival. Cinderella, unexpectedly, is helped by pigeons in sorting the lentils, allowing her to finish early, attend the festival, and meet her prince.

[18] For instance, you believe that <my boss told me that I have to finish the project by 1 December in our last Monday morning meeting > only after you talked to your boss during this meeting and you will have the meta-belief that you believe this since the last Monday morning meeting. Likewise, you would likely associate the stupefying belief that you or a relative is gravely ill with a specific moment in time.

[19] Adopting the externalist perspective on epistemic defeaters does not settle the question whether S has a defeater-deflector from his very own perspective: it merely allows us to uphold the view, in accordance with Moon, that a defeater-deflector is present from t_3 onwards, from some perspective although it is still doubtful whether S, from his own perspective, may draw the same conclusion.

[20] The problem of moral disagreement might also be used to block this reply, cf. Enoch, 2009.

[21] In his tentative answer to MOON'S QUESTION, Moon suggests that the XX-Deflector case is the better case for realists to be in. On my interpretation, this is false: a local defeater, as in the YY-Deflector case, is easier to deal with. But, as I argue in this section, we cannot interpret the reliability challenge in metaethics on the model of a local defeater.

[22] Which presupposes a view according to which all beliefs are the result of cognitive processes, which means that there is no direct, i.e. non-cognitive, perception to form a belief.

[23] Global: all belief-forming faculties. Local: selected *sources* (e.g. information from this-or-that book) or selected belief-forming faculties (e.g. perceptual beliefs).

[24] Even though the idea that moral beliefs might be justified through non-moral beliefs appears to be quite a common one. Shafer-Landau, for instance, discusses and rejects it in relation to Hume's Open Question Argument (Shafer-Landau, 2004, pp. 121ff).

[25] Realists need not accept representationalism, according to which our moral beliefs purport to represent the world, in which case they might have other ways around the reliability challenge (cf. Bogardus, 2016). Most proponents of the form of moral realism relevant in this article, however, do accept a form of representationalism (cf. Enoch, 2011; Shafer-Landau, 2003; Wielenberg, 2014).

[26] Thanks to Hein Duijf and Herman Philipse for helpful discussions of this article and to two anonymous referees for constructive feedback on an earlier version of this article. Special thanks to Liam Deane for perceptive advice and for proofreading the paper.

REFERENCES

Benacerraf, P. (1973). 'Mathematical Truth,' *The Journal of Philosophy* 70(19), pp. 661–679.

Berker, S. (2014). 'Does Evolutionary Psychology Show that Normativity is Mind-Dependent?' in J. D'Arms and D. Jacobson (eds) *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics*. Oxford: Oxford University Press, pp. 215–252.

Bogardus, T. (2016). 'Only All Naturalists Should Worry about Only One Evolutionary Debunking Argument,' *Ethics* 126(3), pp. 636–661.

Boudry, M. and Vlerick, M. (2014). 'Natural Selection Does Care about Truth,' *International Studies in the Philosophy of Science* 28(1), pp. 65–77.

Brink, D. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Brosnan, K. (2011). 'Do the Evolutionary Origins of Our Moral Beliefs Undermine Moral Knowledge?' *Biology and Philosophy* 26(1), pp. 51–64.

de Cruz, H., Boudry, M., De Smedt, J. and Blancke, S. (2011). 'Evolutionary Approaches to Epistemic Justification,' *dialectica* 65(4), pp. 517–535.

Enoch, D. (2009). 'How is Moral Disagreement a Problem for Realism?' *The Journal of Ethics* 13(1), pp. 15–50.

Enoch, D. (2010). 'The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope with It,' *Philosophical Studies* 148(3), pp. 413–438.

Enoch, D. (2011). *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press.

Field, H. H. (1989). *Realism, Mathematics and Modality*. Oxford: Blackwell.

Foley, R. (2001). *Intellectual Trust in Oneself and Others*. Cambridge: Cambridge University Press.

Fraser, B. J. (2014). 'Evolutionary Debunking Arguments and the Reliability of Moral Cognition,' *Philosophical Studies* 168(2), pp. 457–473.

Gibbard, A. (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Grundmann, T. (2014). 'Defeasibility Theory,' in S. Bernecker and D. Pritchard (eds) *The Routledge Companion to Epistemology*. New York: Routledge, pp. 156–166.

Haley, K. and Fessler, D. (2005). 'Nobody's Watching?' *Evolution and Human Behavior* 26(3), pp. 245–256.

Hanson, L. (2016). 'The Real Problem with Evolutionary Debunking Arguments,' *The Philosophical Quarterly*, pp. 1–26. https://doi.org/10.1093/pq/pqw075.

Joyce, R. (2006). *The Evolution of Morality*. Cambridge, MA: MIT Press.

Joyce, R. (2016). 'Evolution, Truth-Tracking, and Moral Scepticism,' in R. Joyce (ed.) *Essays in Moral Skepticism*. Oxford: Oxford University Press, pp. 142–158.

Klenk, M. (2017). 'Old Wine in new Bottles. Evolutionary Debunking Arguments and the Benacerraf-Field Challenge,' *Ethical Theory and Moral Practice*, pp. 1–15. https://doi.org/10.1007/s10677-017-9797-y.

Locke, D. (2014). 'Darwinian Normative Skepticism,' in M. Bergmann and P. Kain (eds) *Challenges to Moral and Religious Belief: Disagreement and Evolution*. Oxford: Oxford University Press, pp. 220–236.

Mogensen, A. (2016). 'Do Evolutionary Debunking Arguments Rest on a Mistake about Evolutionary Explanations?' *Philosophical Studies* 173(7), pp. 1799–1817.

Moon, A. (2012). 'Three Forms of Internalism and the new Evil Demon Problem,' *Episteme* 9(4), pp. 345–360.

Moon, A. (2016). 'Debunking Morality: Lessons from the EAAN Literature,' *Pacific Philosophical Quarterly*, pp. 1–18. https://doi.org/10.1111/papq.12165.

Nickerson, R. (1998). 'Confirmation Bias: A Ubiquitous Phenomenon in Many Guises,' *Review of General Psychology* 2(2), pp. 175–220.

Plantinga, A. (2000). *Warranted Christian Belief*. Oxford: Oxford University Press.

Pollock, J. (1970). 'The Structure of Epistemic Justification,' *American Philosophical Quarterly* 4, pp. 62–78.

Railton, P. (1986). 'Moral Realism,' *The Philosophical Review* 95(2), pp. 163–207.

Ruse, M. (1998). *Taking Darwin Seriously: A Naturalistic Approach to Philosophy*. Amherst, NY: Prometheus Books.

Schacter, D., Guerin, S. and St Jacques, P. (2011). 'Memory Distortion: An Adaptive Perspective,' *Trends Cogn Sci* 15(10), pp. 467–474.

Schechter, J. (2013). 'Could Evolution Explain Our Reliability about Logic?' in T. Gendler and J. Hawthorne (eds) *Oxford Studies in Epistemology, Volume 4*. Oxford: Oxford University Press, pp. 214–250.

Shafer-Landau, R. (2003). *Moral Realism: A Defence*. Oxford: Clarendon Press.

Shafer-Landau, R. (2004). *Whatever Happened to Good and Evil?* New York: Oxford University Press.

Skarsaune, K. (2011). 'Darwin and Moral Realism: Survival of the Iffiest,' *Philosophical Studies* 152(2), pp. 229–243.

Street, S. (2006). 'A Darwinian Dilemma for Realist Theories of Value,' *Philosophical Studies* 127(1), pp. 109–166.

Street, S. (2008). 'Reply to Copp: Naturalism, Normativity, and the Varieties of Realism Worth Worrying about,' *Philosophical Issues* 18(1), pp. 207–228.

Street, S. (2016). 'Objectivity and Truth: You'd Better Rethink It,' in R. Shafer-Landau (ed.) *Oxford Studies in Metaethics: Volume 11*. Oxford: Oxford University Press, pp. 293–334.

Tomasello, M. (2016). *A Natural History of Human Morality*. Cambridge, MA: Harvard University Press.

Vavova, K. (2015). 'Evolutionary Debunking of Moral Realism,' *Philosophy Compass* 10(2), pp. 104–116.

White, R. (2010). 'You Just Believe that because…,' *Philosophical Perspectives* 24(1), pp. 573–615.

Wielenberg, E. (2010). 'On the Evolutionary Debunking of Morality,' *Ethics* 120(3), pp. 441–464.

Wielenberg, E. (2014). *Robust Ethics: The Metaphysics and Epistemology of Godless Normative Realism*. Oxford: Oxford University Press.