The Philosophical Quarterly, *Vol. 55, No. 221* *October 2005*
ISSN 0031–8094

PUTTING EXTROSPECTION TO REST

By Aaron Z. Zimmerman

Jordi Fernández has recently responded to my objection that his 'extrospectionist' account of self-knowledge posits necessary and sufficient conditions for introspective justification which are neither necessary nor sufficient. I show that my criticisms survive his response unscathed.

Jordi Fernández claims that self-knowledge is grounded in 'extrospection'.[1] More often than not, when a subject *S* has good evidence for believing some proposition *p*, *p* is true (I shall call this fact the 'mind to world regularity'). When *S* has good evidence for believing *p*, *S* typically believes *p* (the 'mind to mind regularity'). Fernández exploits both of these putative correlations to argue that when people know that they believe *p*, the very states of mind that justify them in believing *p* will serve to justify them in believing that they believe *p*. Suppose *S* has good evidence *e* for the truth of *p*. The mental state *m* constituted by *S*'s having *e* will ground or justify *S* in believing *p* because of the positive correlation between the existence of *m* and the truth of *p*. Moreover, *m* will justify *S* in believing that he believes *p* because of the positive correlation between *m* and believing *p*.

In reply, I argued that this extrospectionist account posits necessary and sufficient conditions for being justified in believing that one believes *p* which are neither necessary nor sufficient for justification.[2] To argue against their sufficiency, I proposed a dilemma. Suppose that *e* is good evidence for *p*. Either Fernández can say that *S* cannot have *e* without believing *p*, or he can allow that *S* can have *e* without believing *p*. The first horn of the dilemma claims that if the first of these disjuncts holds, *S*'s having *e* provides superfluous grounds for his belief that he believes *p*. Suppose *S* cannot have *e* without believing *p*. Then when *S* has *e* and believes that he believes *p*, this second-order belief must not only be justified but true. But if *S* believes *p*, why cannot *S*'s believing *p* justify *S* in believing that he believes *p*? And if this belief – i.e., the 'truth-maker' for *S*'s introspective belief – does justify him in believing that he has it, what justificatory role is left for the mental state constituted by *S*'s having *e*? *S* might appeal to his having *e* to convince someone else that he believes *p*, but this would not provide an accurate picture of *S*'s grounds. I might appeal to the fact that I am standing in broken glass in order to convince someone that I am in pain, but my belief that I am standing in glass does not ground my

[1] J. Fernández, 'Privileged Access Naturalized', *The Philosophical Quarterly*, 53 (2003), pp. 352–72.
[2] A. Zimmerman, 'Unnatural Access', *The Philosophical Quarterly*, 54 (2004), pp. 435–8.

belief that I am in pain. Instead, I would argue, my belief that I am in pain is directly generated by and grounded in the pain itself.

These points are so obvious that they are hard to argue for. But I need not search for an independent argument here, because (somewhat surprisingly) Fernández's reply adopts the second horn of the dilemma by allowing that *S* can have good evidence for *p* and still fail to believe *p*.[3] If Fernández chooses this route, I argued, he is forced to admit that false second-order introspective beliefs which are unjustified (and so irrational) are actually justified. Suppose Mary has good scientific evidence e_1 for the proposition p_1 that many physical differences between species are the result of natural selection, but that she fails to believe p_1 in the face of e_1. And suppose (as seems hard even to imagine) that Mary nevertheless falsely believes that she believes p_1 simply because she has e_1. Is Mary justified in believing that she believes p_1? No. She considers whether the relevant differences between species can be accounted for by natural selection and either rejects this proposition or decides she must remain agnostic, and yet she concludes that she believes the proposition because of the evidence for its truth in her possession. She is not sufficiently impressed by the evidence to conclude that the evolutionary explanation is true, but she is sufficiently impressed by her possession of the evidence to conclude that she believes the explanation. That just does not sound like the way in which a rational person forms introspective beliefs, and this observation is reinforced by a consequence of the case which is related to Moore's paradox. If Mary can give voice to her beliefs, she will assert 'I believe the evolutionary explanation', while refusing to issue an affirmative answer when someone asks her 'Is the evolutionary explanation true?'. This pattern of response places Mary's rationality in question.

Fernández does not respond by trying to argue that despite these points, Mary's extrospectionally generated belief that she believes p_1 is extrospectionally justified and therefore rational. Instead, he argues that Mary's second-order belief fails to meet one of the necessary conditions the extrospectionist posits for justification. Let m_1 be the mental state constituted by Mary's having e_1. Fernández says that because m_1 is positively correlated with the truth of p_1, m_1 enters into the appropriate mind to world regularity, but because m_1 is not positively correlated with Mary's believing p_1, m_1 does not enter into the appropriate mind to mind regularity. On the extrospectionist account, Fernández argues, *S*'s having evidence for *p* only justifies *S* in believing that he believes *p*, if having evidence of the relevant type is positively correlated with having a belief of the relevant type.

But this response misses the point of the example. There are two ways to show this. First, one might argue that as the case is described, m_1 is in fact positively correlated with believing p_1 because Mary is the exception rather than the rule. Luckily, most people are rational most of the time, so most people who grasp the evidence that the relevant physical differences between species are accounted for by natural selection will indeed believe the evolutionary explanation. Thus one can resist Fernández's claim that the relevant mind to mind regularity fails to obtain in Mary's case, and one can argue on this basis that the extrospectionist is forced into the implausible claim that Mary is justified in falsely believing that she believes p_1.

[3] Fernández, 'Privileged Access Revisited', *The Philosophical Quarterly*, 55 (2005), pp. 102–5.

Of course, Fernández might argue that what is necessary is not a positive correlation between having e_1 and believing p_1, but instead a correlation between *Mary's* having e_1 and *Mary's* believing p_1. What we have here is just an instance of the so-called 'generality problem' for reliabilist accounts of justification.[4] Let M_1 be the relatively 'broad' mental state type *having* e_1, and let B_1 be the similarly broad mental state type *believing* p_1; let M_2 be the narrow mental state type *Mary's having* e_1 and let B_2 be the narrow mental state type *Mary's believing* p_1. When assessing whether the extrospectionist mechanism that generates Mary's second-order belief is reliable, should we see whether instantiating M_1 is positively correlated with instantiating B_1 or, as Fernández proposes, should we see whether instantiating M_2 is positively correlated with instantiating B_2? Suppose we follow Fernández and accept the latter scheme. Then if Mary only entertains e_1 once and only forms a single judgement as to the truth of p_1, there are only two logically possible ways for the correlation between M_2 and B_2 to turn out: perfectly positive or perfectly negative. Requiring that M_2 must be positively correlated with B_2 in a case meeting this description is therefore tantamount to requiring that Mary's belief that she believes p_1 must be true if it is extrospectionally justified. It should be obvious that establishing this requirement would force Fernández back onto the first horn of the dilemma: if *S* believes *p* in every possible case in which *S* is in an evidential state positively correlated with believing *p*, *S*'s believing *p* can always serve as *S*'s grounds for believing that he believes *p*.

This line of argument reveals the second deficiency in Fernández's reply. It is really no good for Fernández to content himself with arguing for the compatibility of the extrospectionist account with Mary's failure to be justified in falsely believing that she believes p_1. A positive argumentative obligation must also be met. Fernández must supply some actual or hypothetical second-order belief which is justified on extrospectionist grounds despite being false. If he cannot do this, he will remain impaled on the first horn of the dilemma: those who do not think that his theory is self-evident will be forced to conclude that every *possible* extrospectionally justified belief is also true and that extrospectionist justification is therefore never needed. We might then alter the case so that it more clearly meets the requirement that Mary's being in the relevant evidential state must be positively correlated with Mary's believing p_1 so that Mary's extrospectionally justified belief meets the reliability criterion when that criterion is interpreted in accordance with Fernández's narrow scheme of individuation. Suppose Mary is an evolutionary biologist until the age of sixty, when she experiences a brief but robust conversion to creationism. As before, she now thinks about the evidence for p_1 and remains unconvinced: she does not believe p_1. She follows the extrospectionist's procedure and is led by the existence of m_1 to the false conclusion that she believes p_1. M_2 is positively correlated with B_2 in this case. But Mary is not justified in believing that she believes p_1 on the basis of m_1, given that she does not believe p_1. This verdict is reinforced by the Moore's-paradox pattern of Mary's linguistic behaviour, and it is not budged at all by the positive correlation of M_2 and B_2. Why not? Because if Mary is at all like everyone else she will not infer that she believes the evolutionary explanation on the ground

[4] See, e.g., R. Feldman, 'Reliability and Justification', *The Monist*, 68 (1985), pp. 159–74.

that she has long had evidence for its truth, provided she does not currently believe the evolutionary explanation. It really does not matter that her having the relevant evidence led her to believe the explanation in the past; nor does it matter that (though she does not now know this) it will again lead her to that belief in the future. When people truly believe that they believe *p*, they are not typically justified by having evidence for *p*. They are justified by believing *p*.[5]

Fernández also tries to resist my argument against the necessity of those conditions for justification posited by the extrospectionist account. My counter-example involved a subject Mary* who has the prejudiced belief that all Es are N, despite never having had any evidence for the truth of this proposition. (I supposed that when her belief is challenged, Mary* just lies, and claims to have known many Es who were N.) Surely Mary* might know and be justified in believing that she believes that all Es are N, despite having no evidence that Es are N, and so lacking justification for believing this proposition. And if she has no evidence that all Es are N, her having evidence for this proposition cannot justify her second-order belief. The extrospectionist is wrong in thinking that one's justified belief that one believes *p* must be grounded in that which grounds one in believing *p*.

Fernández's reply tries to assimilate this scenario to cases of forgotten evidence, which he thinks the extrospectionist account can handle. I shall remain neutral as to whether forgotten evidence poses a problem; I simply insist that as I described the case, Mary* never had any evidence to forget. Fernández does not tell us what needs to be in place if Mary* is to have reasons or evidence for believing a proposition. Perhaps, then, he is assuming that even beliefs drummed in by raw indoctrination are backed by reasons or evidence. (He might, for instance, think that the fact that Mary* has seen on television fictional portrayals of Es as N constitutes, or generates, such evidence or reasons.) Though I doubt the cogency of this view of evidence, I am nevertheless prepared to grant it for the sake of argument and alter the case to one in which a prejudiced neurosurgeon implants the belief that all Es are N (or its realizer) in Mary*'s brain while she sleeps. When Mary* awakes, she believes that all Es are N, and, as before, she defends the belief with lies. Surely this form of belief-fixation is possible: Mary*'s neural state might match that of a more conventional bigot, and the mental state it realizes might play almost the entire causal/explanatory role of a paradigmatic belief, leading Mary* to infer that Es are no good, to avoid Es whenever possible, to treat those Es she encounters with contempt, to assert sincerely that all Es are N, and so on. It seems clear that Mary*'s surgery would instill in her an entirely groundless prejudiced belief without threatening her justification for believing that she has it.

In the end, then, despite Fernández's admirable ingenuity in formulating responses to these counter-examples, they do indeed show that the extrospectionist's conditions are neither sufficient nor necessary for introspective justification.

University of California, Santa Barbara

[5] I defend this view in 'Basic Self-Knowledge: Answering Peacocke's Criticisms of Constitutivism', *Philosophical Studies*, forthcoming.