Neuroethics (2014) 7:75–91
DOI 10.1007/s12152-013-9183-y

ORIGINAL PAPER

Enhancing Moral Conformity and Enhancing Moral Worth

Thomas Douglas

Received: 27 December 2012 / Accepted: 27 February 2013 / Published online: 12 April 2013

Abstract It is plausible that we have moral reasons to become better at conforming to our moral reasons. However, it is not always clear what means to greater moral conformity we should adopt. John Harris has recently argued that we have reason to adopt traditional, deliberative means in preference to means that alter our affective or conative states directly—that is, without engaging our deliberative faculties. One of Harris' concerns about direct means is that they would produce only a superficial kind of moral improvement. Though they might increase our moral conformity, there is some deeper kind of moral improvement that they would fail to produce, or would produce to a lesser degree than more traditional means. I consider whether this concern might be justified by appeal to the concept of moral worth. I assess three attempts to show that, even where they were equally effective at increasing one's moral conformity, direct interventions would be less conducive to moral worth than typical deliberative alternatives. Each of these attempts is inspired by Kant's views on moral worth. Each, I argue, fails.

Keywords Neuroenhancement · Moral enhancement · Moral improvement · Moral worth · Kant

T. Douglas (✉)
Balliol College,
Broad Street,
Oxford OX1 3BJ, UK
e-mail: thomas.douglas@philosophy.ox.ac.uk

T. Douglas
Oxford Uehiro Centre for Practical Ethics,
Faculty of Philosophy, University of Oxford,
Oxford, UK

Morality gives us reasons to do, and not to do, certain things. It may also sometimes give us reasons to do certain things from certain motives, but let us focus, for the moment, on moral reasons to act that are insensitive to one's motives for acting.[1]

Let us say that an agent *conforms to morality* or *morally conforms* to the extent that her conduct coincides with these moral reasons. An agent fully conforms to morality on a given occasion when she performs an act that is at least as well supported by moral reasons as any alternative act, and she fully conforms to morality over a period of time when she performs a series of acts that is at least as well supported by moral reasons as any alternative series.[2]

Most of us would, given a moment's reflection, have little difficulty identifying various ways in which we regularly fail to fully conform to morality. Perhaps we are insufficiently attentive friends. Perhaps we labour under subconscious sexual and racial biases that lead to subtly discriminatory behaviour. Or perhaps we do too little to prevent or correct global problems like environmental destruction and developing-world poverty. We may not think of these moral failures, taken in isolation, as particularly grievous, but we should acknowledge that they can aggregate with devastating effect. Arguably, our failures of moral conformity are, taken together, a driving force behind climate change and global

[1] Of course, performing a given action may itself require having certain motives. For example, whether an item of behaviour that causes death counts as an action of killing will depend in part on whether the person whose behaviour it was intended the death. In stating that I will focus on moral reasons to act that are insensitive to one's motives for acting I mean only insensitive to motives that are not required for one's action to be of the relevant type.

[2] In each case, 'act' should be understood to include inaction.

poverty. It is also increasingly recognised that many of history's greatest atrocities—ranging from the First World War to the Final Solution to the Cultural Revolution—were made possible by the ordinary moral failures of ordinary people [1–3].

It is plausible that we have reasons to correct our moral failures, bringing it about that we better conform to morality.[3] However, this is not to say that we ought to pursue greater moral conformity by any means available. There may be some means to increased moral conformity that we have conclusive moral reasons to avoid, and even among means that are not absolutely ruled out in this way, some means might be better supported by moral reasons than others. There is, in my view, much interesting work to be done in assessing the morality of different possible means to greater moral conformity.[4]

In a recent series of articles, John Harris has begun to do this work [4–6]. Harris has argued that we have reason to adopt traditional, deliberative means of increasing moral conformity in preference to certain more novel means that have recently been discussed by a number of other authors [7–13].[5] He does not precisely delineate the classes of intervention that he favours and disfavours. However, he does identify some members of each class. For example, he explicitly places within the favoured category attempts to increase one's moral conformity though development of a "sophisticated understanding of cause and effect" or through "self-education, wide reading and engagement with the world" [4:104].[6] On the other hand, he raises concerns about a class of interventions that has been defended by Thomas Douglas. Douglas argues that it would sometimes be permissible for individuals to directly influence their emotions—for example, through the use of neurally active drugs—in ways that can be expected to leave them with morally better motives or conduct [9].[7] But Harris objects to interventions that are "targeted on the emotions" in this way [6:1]. He allows that the voluntary use of such interventions to increase one's moral conformity might sometimes be morally permissible or even desirable. But he argues that we nevertheless have reason to *prefer* more traditional, deliberative means to increased moral conformity.[8]

In this article, I assess one concern that might be offered in support of this view. I call this the Superficiality Concern.

The Superficiality Concern

Harris does not unambiguously state the Superficiality Concern, but it can be distilled from a number of asides that he offers while setting out other concerns.[9] The most revealing passages can be found in a discussion of my definition of a class of interventions—emotional moral enhancements—as interventions that (i) "will expectably leave an individual with more moral (*viz.*, morally better) motives or behaviour than she would otherwise have had" and (ii) operate via the direct modulation of emotions [10:3]. The distinctive feature of *emotional* moral enhancements, I claimed that "once the enhancement has been initiated, there is no further need for cognition: emotions are modified directly".[10] Harris objects that "[t]his so-called distinctive feature . . . shows that this concept cannot be moral enhancement properly so called at all". An intervention that operates in this way "is hardly an enhancement, and certainly not one that has much to do with morality" [5:4]. Indeed, he maintains that "the notion of moral behaviour has been attenuated to a vanishing point" once one claims that such behaviour could be produced by directly altering emotions [5:6]; "tinkering with the emotions is not a form of moral enhancement at all. It is more like the threat of punishment: it may make immoral behaviour less likely, but it does not enhance morality" [6:3–4].

One way of reading these passages would see them as an outright denial of the possibility of morally

[3] 'We' refers here to all moral agents who do not already fully conform to morality. This category plausibly includes all mentally competent adult persons that have ever existed.

[4] In this paper I consider only interventions that aim to increase one's own moral conformity. I remain silent on interventions that aim to increase the moral conformity of *others*.

[5] Harris does not specify the nature of these reasons, but I take them to be *pro tanto* moral reasons.

[6] See also Harris [5].

[7] See, for similar arguments, Faust [11] and DeGrazia [13].

[8] Relatedly, Robert Sparrow [Sparrow, Robert. 2013. Better Living Through Chemistry? Unpublished, Sparrow, Robert. 2013. (Im)moral technology? Thought experiments in the future of 'mind control'. Unpublished.] argues that we should prefer political means of improving moral conformity to pharmaceutical or neurotechnological ones. As we shall see, Sparrow's concerns regarding pharmaceutical and neurotechnological means substantially overlap with Harris' concerns regarding the direct modulation of emotions.

[9] These are the concerns that interventions which directly target the emotions would restrict freedom and that attempts at such enhancements would frequently misfire, bringing about morally worse, not better, motives and conduct.

[10] This passage is cited in Harris [5:4] and comes from an unpublished draft version of Douglas [10].

improving motivation or conduct by directly manipulating emotions. However, surely Harris would accept that direct manipulation of emotions could result in at least one kind of moral improvement: it could increase the moral *conformity* of one's conduct.

Note first that the enhancement of moral conformity through directly modulating emotions is nomologically possible—that is to say, it does not violate any laws of nature. Emotions are mental states, mental states are normally taken to be either constitutively or causally dependent on brain states,[11] and there is no law of nature ruling out the direct modulation of the relevant brain states. Thus, it is nomologically possible for direct interventions to alter the emotions. And it is surely also nomologically possible for the alteration of one's emotions to increase the moral conformity of one's conduct.

Perhaps Harris' thought was not that it is nomologically impossible to increase moral conformity through direct emotion-modulation, but simply that this is unlikely to become technologically feasible. However, this suggestion also seems dubious. Consider this case:[12]

> *Andrew* is a doctor working in multi-racial area. He was brought up in a racist environment and emotional responses introduced during his childhood still have a biasing influence on his conduct. For example, they incline him to take more care in treating White patients than Black patients. Andrew is aware of this aspect of his psychology and suspects it to be morally problematic. Hoping to mitigate his bias, he embarks on new programme developed by neuroscientists. He first observes stimuli that elicit racial aversion (such as photos of mixed race couples and civil rights protests) while undergoing high-resolution brain scanning to determine which neural connections mediate the aversion. Those connections are then selectively attenuated via regular sessions of transcranial electrical brain modulation. This programme significantly weakens his disposition to racial aversion and does indeed lead him to treat his Black and White patients more equally.

It is somewhat plausible that an intervention of the kind described here would increase Andrew's moral conformity, and it is not fantastic to suppose that such an intervention might be developed in the future. After all, transcranial electrical brain modulation can already be used to alter rather specific mental abilities such as numerical competence [15] and the ability to deceive others [16].

It seems difficult to deny the possibility of enhancing moral conformity through direct emotional modulation, and this is so whether possibility is understood as nomological possibility or as likely technical feasibility.[13] However, there is a more plausible way of understanding Harris' concern. We could instead take Harris' claim that the direct modulation of emotions could not produce "moral enhancement properly so-called" to be the claim that, although such modulation could increase moral conformity, there is some deeper variety of moral improvement that it would not produce, or, at least, that it would produce to a lesser degree than the more traditional ways of improving moral conformity that Harris favours.[14] Thus, it would result in a kind of moral improvement that is, in one

[11] There are variants of mind–body dualism which deny that mental states are causally or constitutively dependent on brain states. For example, G. W. F. von Leibniz [14] famously subscribed to mind–body parallelism, a version of dualism which takes mental states and bodily states to be causally independent of one another. However, such views are now philosophically obsolete. Contemporary dualists are typically either epiphenomenalists or interactionists, and both of these views allow that all mental phenomena have physical causes.

[12] The case is modified from Douglas [10:2].

[13] It might, however, plausibly be argued that no means of directly modulating emotions that are likely to be developed in the foreseeable future would produce *reliable* moral conformity. Some authors hold that (i) reliable moral conformity can only be achieved through the exercise of moral judgment and (ii) that moral judgment cannot be codified as a simple decision-making procedure. See, for example, Hursthouse [17:230–1]. It might follow that the only reliable way to improve moral conformity is to refine one's noncodifiable moral judgment, and it might seem unlikely that any direct emotion-modulating intervention developed in the foreseeable future could do this (I address this worry in the section "Unreliable Moral Conformity" below). However, even those who take this line would surely accept that these interventions could increase moral conformity in a semi-reliable way, for example, because they induce motivational states that happen to roughly mirror those that would have been produced by the exercise of mature moral judgment.

[14] Harris is most naturally interpreted as claiming that there is some deeper variety of moral improvement that the direct modulation of emotions could not produce *at all*, however, I here attribute to him only the weaker view that such interventions would not produce this deeper variety of moral improvement *to as great a degree* as the more traditional, deliberative interventions that he favours. As we shall see, even this weaker claim is difficult enough to sustain.

respect at least, more superficial than that produced by these more traditional means.

It will be helpful to introduce some terminology here. Call an intervention undergone by some agent a *conformity enhancement* if and only if (i) one of the agent's aims, in undergoing the intervention, is to increase her moral conformity during some extended future time period, and (ii) the intervention succeeds in realising that aim. Harris Superficiality Concern, as I will understand it, maintains that, though all conformity enhancements by definition increase moral conformity, some kinds of conformity enhancement—those that employ certain *direct* means—fail to produce, or fail to produce to the same degree, a deeper kind of moral improvement that is typically produced by traditional, deliberative conformity enhancements.

Similar claims have been made by other authors concerned by certain means of enhancing moral conformity. For example, Fabrice Jotterand [18:8] argues that neurotechnological interventions intended to increase moral conformity are "unlikely to morally enhance people in the true meaning of the word". Similarly, Robert Sparrow [Sparrow, Robert. 2013. Better Living Through Chemistry? Unpublished.], suggests that "while there is indeed evidence that certain pharmaceutical and neuroscientific interventions can alter dispositions and behaviour in ways that we may be inclined to morally evaluate positively, this falls well short of constituting 'moral bioenhancement' in any interesting sense. . . [T]he prospect of making people 'more moral' through pharmaceutical or surgical interventions is slim indeed." He argues further that, whereas commentators in this area have often supposed that "altering behaviour—to prevent someone acting immorally or to ensure that they do the right thing in some particular circumstances—is 'moral enhancement'," this is too quick:

> the use of the sedative gas can prevent someone completing an assault and we would hardly think that this was a case of moral enhancement. At the very least, moral bioenhancement must improve people's motivations.
>
> However, . . . even altering motivation as well as behaviour seems to fall significantly short of enhancing individuals' morality. We are . . . all familiar with drugs that can alter how we feel . . . [A]nyone who has had a few glasses of beer knows that drugs can make us feel love where we would otherwise feel apathy or brave where we would normally be scared. In some circumstances, these chemically influenced emotions may even motivate us to do the right thing. Yet, again, it stretches credulity to call this 'moral enhancement' . . . A stiff shot of whiskey might allow us to summon up the 'courage' required to act morally in some particular instance but it will not succeed in making us 'more moral'.

The Superficiality Concern, as I have outlined it so far, is thus not unique to Harris. However, in discussing that Concern in what follows, I will guided primarily by Harris' discussion of the concern, drawing on other authors only insofar as their worries overlap with his. Accordingly, unless otherwise specified, 'the Superficiality Concern' refers to Harris' variant of the concern.

The Scope of the Superficiality Concern

More on the content of the Superficiality Concern will follow, but first, it will be useful to say something about which conformity enhancements fall within the scope of the Concern, and which fall without it.

It seems clear that Harris would raise the concern in relation to the intervention undergone by Andrew in the case set out above; this case is only slightly modified from one that I offered as an example of an emotional moral enhancement, and Harris does not exclude that case from the scope of his concerns. I take it that Harris is also committed to raising his concerns (including the Superficiality Concern) in relation to this case:

> *Bryony* is a student from a wealthy family. She suspects she ought to do more to help the global poor. She does occasionally do *something* to help, for example, giving small amounts to support famine relief when approached by charities, but most of the time, the world's most unfortunate are far from her thoughts, and when they do cross her mind, she has trouble drumming up the sort of sympathy that might motivate greater sacrifices on her part. In an attempt to remedy this, she sets up her television so that it regularly displays disturbing and graphic images of the effects of poverty, though for such brief periods that she does not consciously recognise them. Nevertheless, through subliminal effects, the images do increase her feelings of sympathy, and these feelings stimulate her to make a large donation to Oxfam.

Unlike most of the putatively problematic conformity enhancements discussed by Harris, Bryony's intervention does not employ biomedical technologies. However, it does manipulate emotions directly, where directness is understood, as by Harris and his interlocutors, as implying that, once the intervention is set in motion, it requires no further engagement of deliberative faculties.[15] This suggests that it would fall within the scope of Harris' concerns.

On the other hand, as we have seen, Harris raises no concerns regarding—and indeed endorses—interventions that increase moral conformity through "self-education". I take it, then, that he would have no problem with the intervention described in this case:

> Like Bryony, *Chloe* is a student who suspects she ought to do more to help the global poor, but has trouble drumming up much sympathy for them. In an attempt to remedy this, she goes to her local library and borrows a number of books containing first-hand accounts of life in poverty. Reading and reflecting on this literature augments her feelings of sympathy, and these feelings stimulate her to make a large donation to Oxfam.[16]

This seems an uncontroversial example of self-education.

What is less clear is where Harris would place conformity enhancements that act directly on mental states, but not on emotions. Such, interventions might instead directly alter desires, intentions, or beliefs. These interventions would presumably not qualify as self-education, since that plausibly implies the alteration of mental states *though deliberation*. But nor are they explicitly mentioned by Harris as among the interventions which raise the Superficiality Concern.

There is, however, some textual support for interpreting the Superficiality Concern to be broader than a concern about only the direct manipulation of emotions. For example, in discussing his concerns about the direct modulation of emotions, Harris frequently adverts to the thought that these "bypass" moral reasoning, or moral reflection, or the exercise of moral judgment [5:2,4–5; 20:E183]. Presumably the thought is that these interventions are used to bring about mental transformations of a sort that could otherwise be achieved through these forms of moral deliberation. But the kinds of mental transformations typically induced by moral deliberation include not just changes in emotional states, but also, at the very least, changes in conative states, such as desires and intentions.

In addition, Harris elsewhere characterises his concern as attaching to interventions that operate "directly on the mainsprings of action" [5:2]. This suggests that his concern is with the direct modulation of any motivating mental states, and again, these include conative states as well as affective ones.

In what follows, I will assume that Harris would raise the Superficiality Concern in relation to all conformity enhancements that operate by directly altering affective or conative states.[17] And I take it that an intervention directly modulates an affective or conative state just in case, once the intervention has been initiated, it alters that state without requiring the exercise of deliberative faculties. I will refer to conformity enhancements that meet these conditions as *brute* conformity enhancements and will take the interventions undergone by Andrew and Bryony above to be examples of such enhancements.

Brute conformity enhancements can be contrasted with what I will call deliberative conformity enhancements: conformity enhancements that consist in moral deliberation. These conformity enhancements might involve moral reasoning, introspective reflection on one's moral failures, or calm moral discussion with others. I take it that Chloe's intervention is a deliberative conformity enhancement. Harris' Superficiality Concern, as I will understand it, is that there is some important variety of moral improvement that brute

[15] Harris does, at one point, suggest that environmental manipulations might be immune to the concerns he raises where the decision to undergo such an intervention is itself motived by deliberation or, as he puts it here, is the product of a "self-conscious strategy" [6:2]. However, it is difficult to see how he can consistently take this view, since he raises his concerns regarding biomedical interventions to manipulate emotions even where the decision to undergo the intervention is the result of deliberation.

Harris also explicitly excludes from the scope of his concerns a case that is rather like the case of Bryony but which involves conscious processing of images of poverty. However, he excludes this case on the grounds that viewing and reflecting on the images counts as a kind of conscious deliberation [6:2–3]. It is difficult to see how he could say the same about the case of Bryony, where the images exert their effect subliminally.

[16] Some might object to Chloe's intervention on the ground that she appears to be driven by a desire to become more moral, a motive that some have found problematic. See, for discussion, Sorensen [19]. However, Harris alludes to no such concern, so I see no reason to attribute it to him.

[17] I remain agnostic on whether normative judgments are or comprise conative or affective states, and I allow that, if they do not, then they do not fall within the scope of Harris' concern.

conformity enhancements fail to produce, or produce only to a lesser degree than typical deliberative conformity enhancements.

It is not clear whether Harris would extend this concern to all brute conformity enhancements that might *in principle* be developed in the future—because they are nomologically possible—or only to those that might plausibly be developed within some restricted time frame. Clearly, his claim would be more difficult to sustain if it were interpreted in the former way. I thus opt for the latter interpretation in order to present his concern in the most plausible light. I henceforth take Harris to be raising the Superficiality Concern regarding all brute conformity enhancements of a sort that might plausibly be developed within the medium term future—the next 50 years or so. Unless otherwise specified, 'brute conformity enhancements' refers only to these enhancements.

The Superficiality Concern and Moral Worth

In alleging that brute conformity enhancements are superficial, one is alleging that these enhancements fail to induce (as much of) some deeper kind of moral improvement. But which deeper kind of moral improvement, exactly? Possible candidates for deeper moral improvements might include increases in the moral virtue, moral responsibility, moral understanding, moral knowledge and perhaps even moral status of the agent, as well as increases in the moral virtue and moral worth of the agent's conduct. In what follows I focus on the last of these; I flesh out the Superficiality Concern as a concern that brute conformity enhancements are less conducive to morally worthy conduct than typical deliberative enhancements. (I believe, however, that much of what I will say could be re-framed in terms of moral virtue or moral responsibility without substantially affecting the arguments.)

The distinction between conforming to morality and acting in a way that has moral worth has been a commonplace since Kant. To say that an action has 'moral worth' is, in standard philosophical usage, to say that it reflects well, morally, on the agent—that the agent merits moral praise for having done that act.[18] It is possible for two acts to accord equally well with moral reasons, yet for one to have greater moral worth than the other. Nomy Arpaly gives the example of two people who donate to Oxfam. The two donate the same amount of money, but one donates to improve the state of the world, while the other does so merely because her accountant advises it [22:69]. It is plausible that the actions of the two donors conform equally well to morality, but that the first agent's action has greater moral worth.

This sort of case opens the door to the possibility that different conformity enhancements might be equally effective in increasing moral conformity yet have different effects on moral worth. This in turn raises the possibility of fleshing out the Superficiality Concern as follows:

> (*The Moral Worth Claim*) For all brute conformity enhancements likely to be developed in the medium-term future, whenever an agent has a choice between pursuing that conformity enhancement or achieving the same increment in moral conformity via a typical deliberative conformity enhancement, adopting the brute conformity enhancement will result in less morally worthy conduct.

To avoid repeatedly stating this rather cumbersome claim, I will sometimes paraphrase it as follows: brute conformity enhancements are less conducive to moral worth—or confer less moral worth on the agent's subsequent conduct—than typical deliberative conformity enhancements.

Harris does not himself explicitly flesh out his Superficiality Concern in the way that the Moral Worth Claim fleshes it out. However, this way of spelling out the Concern does sit well with some of what he says in support of it. For example, Harris notes, in defence of his Superficiality Concern, that

> [o]ne can accidentally discover something of scientific importance, but one cannot be scientific, one cannot do science, accidentally. Doing science is a deliberative and disciplined process. It involves, for example, doing things like formulating and testing a hypothesis and looking for disconfirmatory evidence as well as for confirmatory evidence . . . Being moral is like being scientific [5:6].

Harris suggests here that a brute intervention could not help one to 'be moral' in part because it could at best

[18] Note that, although this is a standard way of understanding moral worth, it is not the only way in which it has been understood in recent philosophical literature. For an example of an alternative understanding, see, for example, Johnson [21].

lead one to do something of moral importance *accidentally*. He does not say exactly what he means by 'being moral', but his worry would make perfect sense if he were equating 'being moral' with 'acting in a morally worthy way', for it has often been thought that moral worth requires non-accidental moral conformity. While one can accidentally conform to morality, if one does, one's conduct will lack moral worth.[19]

The Moral Worth Claim also sits well with one of Harris' other concerns about brute conformity enhancements. Harris is concerned that brute conformity enhancements, or at least certain among them, might diminish or restrict the freedom to do wrong, and perhaps one reason why the freedom to do wrong is valuable is that it enables *rightful* action to have moral worth. If we were not free to do wrong, then arguably nothing we did would have moral worth.

Robert Sparrow can be interpreted as appealing more directly to the concept of moral worth to support his variants of the Superficiality Concern. He claims, for example, that acting morally

> requires that agents should respond in the right way to counterfactuals: if we praise someone for helping another person who is in need, our assessment that their action is morally admirable rests upon the thought that they should not have been motivated to help them in the same way if the other person were not in need. It is difficult to see how any pharmaceutical could cause us to have the appropriate beliefs about what moral action would consist in, not only in the current circumstances that we face but also in others that are both relevantly similar and dissimilar. It would be a good drug, indeed, that made us feel love only for what is worthy of love and brave only in the service of a just cause [Sparrow, Robert. 2013. Better Living Through Chemistry? Unpublished].

Sparrow's use of the terms 'praise' and 'morally admirable' in this passage strongly suggests that, by acting morally, he means acting in a morally worthy way.

It may be reasonable, then, to read Harris as implicitly endorsing the Moral Worth Claim, and Sparrow as explicitly endorsing it, or at least something close to it. Moreover, even if these authors do not endorse the Claim, it strikes me as among the more plausible ways of spelling out Harris' Superficiality Concern, and indeed of supporting his view that we have reason to adopt deliberative conformity enhancements in preference to brute ones. Thus, it seems worth considering whether the Moral Worth Claim is correct. In what follows, I will assess three attempts to show that it is.

Throughout, I will simply grant that the Moral Worth Claim would indeed, if correct, support the view that Harris' ultimately wishes to defend; I take this to be the view that we would have reason to adopt a typical deliberative conformity enhancement in preference to any alternative brute conformity enhancement that might plausibly be developed in the medium-term future. In fact, it is not obvious that the Moral Worth Claim does support this view, for it is not obvious that we have any reason to promote moral worth in our own future conduct. Morally worthy conduct is conduct that merits praise; it is not clear that it also merits promotion.[20] However, some have argued that we do have reasons to promote moral worth [e.g., 25:15–17; 26], and for the sake of argument, I shall assume that they are correct.

This assumption places some constraints on what will qualify as an adequate defence of the Moral Worth Claim. Defences will need to be consistent with this claim bearing positively on Harris' view about the preferability of deliberative conformity enhancements. For example, a defence should not establish the Moral Worth Claim at the price of conceding that we have no reason at all to promote moral worth in our future conduct. Moreover, since our motive for assessing the Moral Worth Claim derives from its putative support for Harris' view that we should prefer deliberative conformity enhancements, it is natural to require also that any adequate defence of the Moral Worth Claim would allow this claim to play *an interesting role* in supporting

[19] I discuss this view further in the section "Unreliable Moral Conformity" below.

[20] For claims that moral worth does not merit promotion, see, for example, work by Richard Henson [23] and Allen Wood [24:30; Wood, Allen. 2013. Moral worth, moral merit, and acting from duty. Unpublished]. Allen Wood [Wood, Allen. 2013. Moral worth, moral merit, and acting from duty. Unpublished] puts the point particularly forcefully, claiming that "if a moral agent is dedicated to a meritorious end . . . —for instance, relieving the suffering of many people—then she will naturally care much more about this end than she does whether some of her actions taken toward it have moral worth because they are done from duty. Indeed, Kant's theory does not justify the agent's concern with this at all, unless the case is one where she will fail to act in conformity with duty unless she acts from duty, and then it is dutifulness itself, not action from duty or moral worth, that matters to the agent.

Discussions of the beginning of the Groundwork . . . often treat moral worth as something agents have reason to want their actions to have. But . . . I think this is a thoroughly misguided thought."

Harris' view. For example, a defence should not establish the Moral Worth Claim while also establishing that the Claim provides only exceptionally weak support for Harris' view. I will expect attempts to defend the Moral Worth Claim to be consistent with this Claim bearing both positively and interestingly on Harris' view that we should prefer typical deliberative conformity enhancements to brute alternatives.

Acting from the Right Motives

Why might the conduct produced by a brute conformity enhancement have less moral worth than the conduct produced by a deliberative conformity enhancement even where the conduct conforms equally well to morality? The obvious place to begin the search for an answer to this question is with Kant's idea that to have moral worth, an act must be done for the right reason or from the right motive. There is little agreement about what sorts of motives are the right ones, but on one prevalent view, Kant's own view, the action must be done from the motive of duty. Kant held that "if any action is to be morally good, it is not enough that it should *conform* to the moral law—it must also be done *for the sake of the moral law*" [27:4:390].[21]

The view that morally worthy actions must be done from the motive of duty—henceforth the 'Kantian view'—has often been taken to support the view that moral worth requires deliberation. It is perhaps not obvious that acting from the motive of duty must involve moral reasoning or any other deliberative process. However, on one standard Kantian position, it must; an agent acting from the motive of duty deliberates about what morality requires or recommends. For example, on Barbara Herman's early interpretation of Kant, "[f]or an action to have moral worth, moral considerations must determine how the agent *conceives* of his action (he *understands* his action to be what morality requires), and this conception of his action must then determine what he does" [28:375, my italics]. We should not ascribe moral worth to the conduct of a "man of sympathetic temper . . . whose helpful actions . . . are motivated by his natural response to the plight of others". Why should we not ascribe moral worth in this case? Because this man "acts because he is, literally, moved by others' distress. There need be no moral component in his *conception* of what he does" [28:376–7, my italics]. Note the central role given here to the deliberative concepts of 'conceiving' and 'understanding'. On Herman's view, it will plainly not be enough, for one to act from the motive of duty, that one acts on impulses or inclinations that are aligned with duty. One must *think* about one's duty.

This might be thought to explain why brute conformity enhancements are less conducive to moral worth than typical deliberative ones. After all, the conduct produced by brute conformity enhancements was arrived at in part through non-deliberative means. For example, the decision by Bryony, the apathetic student, to make a large donation to Oxfam was arrived at in part through subconscious processes caused by her subliminal imagery programme.

There is a difficulty with this explanation, however. Some brute conformity enhancements—including ones that might plausibly become technologically feasible in the medium-term future—might operate precisely by facilitating the sort of deliberation that the Kantian, as we are understanding her, takes to be necessary for moral worth. Our earlier example of Andrew, the biased doctor, might, depending on how the details are filled out, be just such a case. The brain modulation programme that attenuates Andrew's racial aversion may help to promote moral conformity precisely because it removes one impediment to the sort of sound moral deliberation that the Kantian values.[22] Even in the case of Bryony, the apathetic student who embarks on a programme of subliminal imagery, it seems possible that the brute conformity enhancement operates by promoting sound moral reasoning. Perhaps by increasing her feelings of sympathy for strangers, Bryony's subliminal imagery programme stimulates her to engage in Kantian-style moral reasoning about how to respond to global poverty. If Andrew's and Bryony's interventions operate as I have just suggested they might, there seems no reason to deny, on the basis of the Kantian view, that their subsequent actions have moral worth. True, brute processes played a role in bringing about these actions. But the proximate aetiology of the agent's action was entirely deliberative in each case. This, plausibly, is all that is necessary to act from the motive of duty. After all,

[21] Volume and page numbers are for the Prussian Academy edition of Kant's collected works. Italics in the original.

[22] Indeed, Douglas' initial proposal was that moral enhancements might operate by mitigating emotions that serve has barriers to good motivation on any plausible account of good motivation (including a broadly Kantian one).

that motive is standardly (though not, as we shall see, universally) taken to be a *proximate* cause of action.[23]

Nevertheless, Harris would, I take it, object to these interventions. Harris appears to regard the Superficiality Concern as most serious in cases where an intervention directly alters the agent's affective or conative states, *and* those alterations in turn directly affect the agent's conduct, without the need for deliberation [5:2–5; 6:4]. Thus, the sorts of cases I have just discussed would perhaps not attract his *most serious* censure. In these cases, changes to affective and conative states affect conduct only indirectly, by facilitating good deliberation. However, these cases nevertheless fall within the scope of his Superficiality Concern. As we have seen, Harris presents his Superficiality Concern as attaching to all conformity enhancements that directly alter emotions, and he can be naturally interpreted as raising it also in regard to those that directly alter conative states. He does not restrict his Superficiality Concern to the subset of these interventions in which alterations to affective and conative states themselves directly influence action.[24] Moreover, if he did restrict the scope of the Concern in this way, he would render it dialectically uninteresting, since none of the authors targeted by Harris have defended conformity enhancements that directly modulate *action*.[25]

Bypassing Deliberation

Even if we accept that (i) morally worthy actions must be done from the motive of duty and (ii) acting from the motive of duty requires deliberation, it seems possible that brute conformity enhancements could be highly conducive to moral worth: they could bring it about that one acts from the (necessarily deliberative) motive of duty. Thus, though (i) and (ii) would plausibly support a restricted variant of the Moral Worth Claim—one that applies only to brute conformity enhancements that directly influence action—they do not support that claim in its original, more general form. The defender of that claim will need to establish that, even where brute conformity enhancements influence action by enabling sound moral deliberation, they fail to be as conducive to moral worth as typical deliberative conformity enhancements.

Although the Kantian view outlined above, as standardly interpreted, does not directly support this position, some ideas that have been thought to underpin that view may support it. One of these is the idea that the causal history or aetiology of an action matters in determining its moral worth. In respect of its proximate aetiology, the conduct produced by brute conformity enhancements might well be beyond reproach, meeting the Kantian requirement that morally worthy actions be done from the (deliberative) motive of duty. However, perhaps there is a problem further back in the aetiology of the conduct. Indeed, it has been argued that Kant himself should be understood as being more focussed on distal motivation than my discussion in the previous section implies [Wood, Allen. 2013. Moral worth, moral merit, and acting from duty. Unpublished]. On this interpretation, the distal aetiology of an action can influence its moral worth.

What might be wrong with the distal aetiology of the conduct produced by brute conformity enhancements? One suggestion would be that the conduct does not *originate* in the deliberation of the agent. This would, I think, be a somewhat promising way of objecting to the imposition of brute conformity enhancements on *others*. Where *A* imposes a brute conformity enhancement on *B*, *B*'s subsequent conduct might be thought to originate not in *B*'s deliberation, but in *A*'s, and this might be thought to detract from its moral worth. However, as the basis for a general worry about brute conformity enhancements, the suggestion seems unpromising. Though it is not at all clear how we should understand the origin of an item of conduct, on any plausible characterisation, it seems that the conduct induced by brute conformity enhancements *could* originate in the deliberation of the agent. This

[23] In more recent work, Barbara Herman [29] has considered the possibility that the 'motive of duty' might be understood not as a proximate cause of an action, but as something that is dispersed among various other causes of action, both proximate and distal. I consider the relevance of distal motives to moral worth in the next section.

[24] Harris does exclude, from the scope of his concerns, conformity enhancements that operate via the biomedical enhancement of cognitive capacity [5:9; 6:4]. This might lead one to suppose that he would have no objection to conformity enhancements that directly manipulate mental states *and thereby facilitate good deliberation*. However, this interpretation is difficult to square with the fact that Harris takes, as the target for his concerns, the kinds of interventions defended in Douglas [9], where I explicitly focus on interventions that alter emotions and thereby, inter alia, facilitate better deliberation.

[25] Indeed, Douglas' initial proposal was that 'moral enhancements' might work by attenuating emotions that serve as barriers to good motivation on any plausible account of good motivation, including a broadly Kantian one which takes sound moral reasoning to be the only good motive [9].

is because the decision to engage in the brute enhancement may itself be arrived at through deliberation.

A more promising suggestion would be that the problem with conduct induced by brute conformity enhancements is that its aetiology was not deliberative all the way down. That is to say, some steps in the aetiology of that conduct that could in principle have been accomplished through deliberation are bypassed—they are taken through non-deliberative means.[26] Andrew, the biased doctor, attenuates his racial aversion, and mitigates his biased conduct, through a programme of electrical brain modulation. He might, perhaps, have achieved the same attenuation of racial aversion through deliberation, for example, by reflecting on his racial aversions, and perhaps by reading about their likely effects. But Andrew did not take these deliberative steps—he bypassed them. He does leave himself with some deliberative work to do. Following the electrical brain modulation programme, he must still deliberate about, for example, how to treat a particular patient on a particular occasion. Thus, he has not *entirely* bypassed deliberative processes. However, in attenuating his racial aversion via the use of pharmaceuticals, he has used brute means to make some progress towards moral conformity in the sense that he has strengthened his disposition to morally conform. This is progress that he could in principle have made deliberatively, for example, through introspective reflection or reflective engagement with literature.

Avoiding Effort

Why should bypassing deliberation limit the moral worth of one's subsequent conduct? One answer is suggested by reflecting on the following pair of cases:

> Compared to his peers, *David* conducts himself in a way that accords well with the moral reasons that apply to him. Indeed, he finds it easy to morally conform since he was brought up in a nurturing family where responsibility and moral sensitivity were encouraged and his role models seldom exhibited or endorsed objectionable moral attitudes. He also lives in a society that has internalised few problematic norms and encourages moral reflection and open moral discussion. It is not that he *automatically* does what morality requires; he frequently has to deliberate about what to do. But his deliberation is seldom biased or disrupted by powerful impulses or misguided social pressures, and sound deliberation is facilitated by the ease with which he is able to imagine the consequences of his actions and empathise with those he affects.
>
> Unlike David, *Felix* was raised in a dysfunctional family where violence was openly encouraged, bigoted attitudes were routinely expressed and endorsed, and moral sensitivity was viewed as a sign of weakness. He also lives in a society that has embraced an objectionable moral code, so social pressures militate strongly in favour of moral nonconformity. Nevertheless, Felix frequently engages in moral deliberation and, despite the distorting influence of deeply engrained emotions and the consistently negative influence of those around him, he is able to conform well to morality—as well, in fact, as David.

Some would, I think, intuit that, at least in one respect, Felix's conduct has greater moral worth than David's. One natural way of accounting for this intuitive response would be to hold that expending moral effort—effort to morally conform—confers moral worth on the resulting actions. Felix's actions have (in one respect) greater moral worth than David's because he expended, on average, greater moral effort in bringing about that conduct.

Others, I suspect, would reject the intuitive response that I have just noted, but in any case, the view that expending moral effort confers moral worth on one's actions has enough philosophical support that it ought to be taken seriously.[27] Moreover, it might seem that, if that view is correct, it will lend support to the Moral Worth Claim. Deliberation is typically an effortful process, but perhaps directly altering one's affective or conative states is not. It might be thought

[26] Of course, it is plausible that no conduct is deliberative 'all the way down' in the strong sense of being motivated wholly through deliberation, and not at all through nondeliberative channels. The relevant point here is that the actions that result from brute conformity enhancements might be thought less deliberative than they could have been.

[27] See, for example, Sorensen [30]. Kant himself also offers some support for this view in *The Metaphysics of Morals*, writing that "[t]he greater the natural obstacles (of sensibility) . . . so much the more merit is to be accounted for a good deed" [31:6:228]. Volume and page numbers are for the Prussian Academy edition of Kant's collected works.

that, when an individual could achieve a given increment in moral conformity though either undergoing a brute conformity enhancement or engaging in deliberation, the brute conformity enhancement will invariably involve exerting less effort.[28] If this is so, then we might have good grounds to suppose that brute conformity enhancements are less conducive to moral worth than typical deliberative conformity enhancements. Undergoing a brute conformity enhancement will make one more like David and less like Felix.

One problem faced by this line or argument is that, intuitively, actions can possess a very high degree of moral worth even if they are relatively effortless.[29] David's moral conformity required less moral effort than Felix's, and perhaps this makes David's conduct somewhat less worthy, at least in one respect. Nevertheless, it is plausible that David's actions frequently possess a very high degree of moral worth. Even those who defend the view that moral effort confers moral worth allow that relatively effortless actions can also be highly morally worthy, because moral effort is not the *only* ground of moral worth. Indeed, on one view, high levels of moral worth are attained at both ends of the spectrum of moral effort, namely, in (i) cases where heroic levels of moral effort are exerted, *and* (ii) cases in which very little moral effort is exerted, because little is needed: the agent's motives are so well-aligned with her moral reasons [30]. If this view is correct, then actions produced through brute conformity enhancements might possess a high degree of moral worth even if the brute conformity enhancement significantly diminishes the amount of effort necessary to perform those actions.

It might be objected at this point that, though actions produced by brute conformity enhancements could be highly morally worthy, they are nevertheless *less* morally worthy than comparable actions produced through effortful deliberative moral remedies. For example, it might be held that, though there are routes to moral worth besides the exertion of moral effort, *ceteris paribus,* more moral effort results in greater moral worth. Thus, suppose *Ervin* started out with a psychology rather like Felix's, so found it difficult to conform to morality, but then underwent a conformity enhancement which left him with a psychology like David's. Ervin's later actions, like David's, might possess a high degree of moral worth. Yet it might be thought that they would have been even worthier had he achieved the same transformation (to a David-like psychological set-up) through effortful, deliberative means rather than effortless, brute ones.

I am not convinced that this response can succeed. An Aristotelian might, for example, maintain that, even if exerting effort can confer some degree of moral worth on one's actions, the morally *worthiest* actions are those whose aetiology features no significant moral effort. I cannot defend that view here. But it is, I think, a reasonable one. Moreover, if this view is correct, then, even if adopting a brute conformity enhancement in preference to a deliberative alternative avoids the exertion of moral effort, it may have no negative influence on moral worth. However, I will not pursue this thought. Instead, I turn to what is, I think, an even more serious problem for the effort-based argument for the Moral Worth Claim.

The problems that I have been discussing stem from the fact that high degrees of moral effort are clearly not *necessary* for an action to have high moral worth, and may not even be necessary for the action to have maximal moral worth. But there is another problem: exerting moral effort does not always confer moral worth, that is to say, it is not *sufficient* for it.

Thus, recall Chloe, who achieved an increment in moral conformity though reading and reflecting on first-hand reports of life in poverty. Suppose that Chloe could also have achieved this increment through pure introspective reflection—that is, *without* the aid of literature. These options, both deliberative, would have been equally effective in increasing moral conformity, but suppose that the purely introspective route to increased conformity would have required greater effort. If this additional effort would have been gratuitous—that is to say, if, leaving aside considerations of moral worth, Chloe had no more reason to adopt the more effortful route than the less effortful one—then it seems very doubtful whether exerting that additional effort would have conferred any moral worth on her subsequent conduct. It would surely be surprising if

[28] It is not at all obvious that brute conformity enhancements would always involve less effort than alternative deliberative ones. One might imagine, for example, that some individuals would have to exert rather great moral effort in order to undergo a brute conformity enhancement because, say, they feel repulsed by the thought of directly influencing their conative of affective states.

[29] I thank an anonymous reviewer for pressing me to discuss this point.

moral worth could be bought through the exertion of effort that there is no moral reason to exert.

Consideration of this case undermines the simple view that moral effort confers moral worth since it suggests that gratuitous moral effort does not. However, it also suggests a more plausible view—namely, the view that *nongratuitous* moral effort confers moral worth. (Note that this modified view might also seem able to account for the intuitive responses that I speculated some might have to the *David* and *Felix* cases, for we might well suspect that Felix has exerted more nongratuitous effort to morally conform than has David.)

Importantly, however, it seems doubtful whether this modified view will be helpful to anyone who wishes to appeal to the Moral Worth Claim in order to defend Harris' view—namely the view we have reason to prefer deliberative conformity enhancements to brute alternatives.

To see the problem, suppose that an agent can bring about a given increment in moral conformity either through a (more effortful) deliberative route or a (less effortful) brute intervention. Suppose initially that an agent has more reason, leaving aside considerations of moral worth, to adopt the deliberative route than the brute alternative. In that case, the additional effort entailed by the deliberative route will be nongratuitous and will thus contribute to the moral worth of the agent's subsequent conduct.

But now suppose instead that, leaving aside considerations of moral worth, the agent does not have more reason to adopt the deliberative route. In this second variant of the case, if the agent opts for the more effortful deliberative route, she will simply be exerting gratuitous effort, and this will confer no moral worth on her subsequent conduct.

Thus we see that adopting a more effortful deliberative conformity enhancement in preference to a less effortful brute alternative confers greater moral worth on one's subsequent conduct only if one already has most reason to prefer that option. Insofar as exerting nongratuitous effort is what matters for moral worth, considerations of moral worth will at most add a supplementary reason to prefer the deliberative route. They will never give an agent most reason to prefer the deliberative route when she would not already have had most reason to do so.

This effectively relegates the Moral Worth Claim to an accessory role in justifying Harris' view that we have reason to adopt deliberative conformity enhancements in preference to brute ones. An appeal to effort may support the Moral Worth Claim *if* we assume that there would be more reason to engage in a typical deliberative conformity enhancement than any brute alternative. This assumption guarantees that the additional effort associated with the deliberative route is nongratuitous. However, if we assume this, then there is no need to appeal to the Moral Worth Claim in order to justify Harris view; we already have good grounds to accept it. I take it, then, that a defender of Harris' view would not want to defend the Moral Worth Claim in this way. Doing so would deprive the claim of most of its interest as a basis for preferring deliberative conformity enhancements.

Unreliable Moral Conformity

An alternative defence of the Moral Worth Claim would appeal to the thought that brute conformity enhancements would produce less reliable moral conformity than typical deliberative conformity enhancements.

According to the Kantian view described earlier, an action must be done from the motive of duty if it is to have moral worth. One of the thoughts that has often been taken to support this view is the thought that actions which accidentally conform to morality lack moral worth. As Barbara Herman puts it, the action's moral conformity must be "the nonaccidental effect of the agent's concern"—"we need to know that it was no accident that the agent acted as duty required" [28:366,368].

In Kant's famous example, a shopkeeper charges his customers fair prices—in conformity with morality—but does so solely in order to maximise his own profit. Kant maintains that the shopkeeper's actions lack moral worth. Herman explains why as follows:

> the moral fault with the profit motive is that it is unreliable. When it leads to dutiful actions, it does so for circumstantial reasons . . . This example suggests the need for a motive that will guarantee that the right action will be done [28:363].

Turning to Kant's 'sympathetic man', whose natural inclinations lead him to help those in distress,

Herman again cites the accidental nature of the man's moral conformity in explaining why his conduct lacks moral worth:

> He acts because he is, literally, moved by others' distress. There need be no moral component in his conception of what he does. Therefore, nothing in what motivates him would prevent his acting in a morally impermissible way if that were helpful to others, and it is to be regarded as a bit of good luck that he happens to have the inclination to act as morality requires [28:377].

This idea—that the moral conformity of an action cannot be an accident if the action is to have moral worth—has been found plausible by many, including some who reject the Kantian view that morally worthy actions must be done from the motive of duty [e.g. 32:206]. There are different ways in which we might make sense of the idea of nonaccidental or reliable moral conformity, but on one standard account, for an agent's moral conformity to be reliable, it must be the case that the agent would also have morally conformed in possible worlds apart from the actual one.[30] Determining precisely which possible worlds bear on the reliability of an agent's moral conformity is a complicated matter, and I cannot address it here. I simply assume that it is in principle possible to provide an account of reliable conformity in terms of counterfactuals, and also that it is in principle possible, by enumerating the number of relevant counterfactual scenarios in which the agent would have conformed, to use this account to generate a measure of the reliability of an agent's moral conformity.

These thoughts on the *reliability* of moral conformity are relevant to our assessment of the Moral Worth Objection to brute conformity enhancements, for it might plausibly be thought that brute conformity enhancements would invariably fail to produce such reliable moral conformity, understood in counterfactual terms, as can be achieved through deliberation. Deliberative conformity enhancements frequently work by enhancing an agent's moral knowledge, moral understanding or moral judgment—henceforth, collectively, her moral-epistemic resources. Through deliberation, the agent comes to know that she has certain moral reasons, acquires an understanding of why she has certain moral reasons, or becomes better at assessing and weighing moral reasons. These moral-epistemic resources are all-purpose tools that help her to morally conform in many circumstances. If an agent knows what moral reasons there are and understands why, and if she is good at assessing and weighing these moral reasons, then, provided she is also somewhat disposed to act in accordance with her moral judgments, she will be well-placed to do what she has most moral reason to do in almost any circumstance. Admittedly, moral-epistemic resources do not translate into moral conformity in all circumstances. One can know what moral reasons there are, or even correctly judge what morality requires of one in a particular case, yet fail to morally conform, for example, due to weakness of will. Still, an agent who possesses substantial moral-epistemic resources will generally be disposed to morally conform across a wide range of possible circumstances.[31]

On the other hand, it might be thought that *brute* conformity enhancements would not normally operate by enhancing the agent's moral-epistemic resources. Rather, they would typically work by removing some relatively straightforward affective or conative obstacle to moral conformity. The most obvious examples of such obstacles might include a tendency toward impulsive violence, strongly xenophobic sentiments or a disinclination to feel sympathy for strangers. But note that these are not *universal* barriers to moral conformity. They obstruct moral conformity only in certain circumstances. Consider the tendency towards impulsive violence. While this may often be a barrier to moral conformity, there are circumstances in which it might instead be conducive to such conformity; for example, when one is fighting a just war, or perhaps when one is confronted with one person assaulting another on the street. Consider alternatively a

[30] On a variant of this view, what matters is whether the agent would have morally conformed *for the right reasons* in other possible worlds. I do not explicitly discuss this view in what follows. However, this view could be substituted for the one I do discuss without substantially affecting my arguments.

[31] I am suggesting, here, that possession of the moral-epistemic resources might contribute to reliable moral conformity and thereby be relevant to moral worth. Some would argue that the moral-epistemic resources are relevant to moral worth as well or instead in a more direct way. For example, Alison Hills has recently argued that acting on the basis of moral understanding contributes directly to the moral worth of one's action [33]. I will not explicitly address the view that moral-epistemic resources are directly relevant to moral worth, however what I say below regarding the view that they are *indirectly* relevant applies equally to the view that they are directly relevant.

tendency to be indifferent to the suffering of others. Though this might often be a barrier to moral conformity, there are circumstances in which it would be conducive to it: these may include those circumstances in which one is an emergency medic surrounded by severe pain and suffering or a judge charged with impartially weighing the claims of plaintiff and defendant. In these settings, a degree of indifference to the suffering of strangers may lead to greater moral conformity.

Though brute interventions which attenuate the tendency toward impulsive violence or lessen indifference to the suffering of strangers might *in fact* increase moral conformity, the moral conformity they produce will be highly contingent on what circumstances obtain. This is one important respect in which any moral conformity produced by such interventions is *unreliable*—perhaps more unreliable, typically, than that produced by deliberative interventions.

There is also another. Whether tendencies towards impulsive violence and indifference to the suffering of strangers impede moral conformity depends on the degree to which those tendencies are present. For example, though a strong tendency towards impulsive violence is unlikely to be conducive to moral conformity, a milder tendency of the same kind may well be conducive to it, for example, because it helps to prevent excessively submissive conduct. Similarly, though thoroughgoing indifference to the suffering of strangers may well impede moral conformity, *some* tendency to ignore the suffering of strangers is presumably conducive to moral conformity; an individual overwhelmed by sympathetic responses to the suffering of others is unlikely to conform well to morality.

These thoughts suggest that there is a further respect in which brute conformity enhancements may produce less reliable moral conformity than deliberative ones. The moral conformity produced by these interventions may be more contingent on the degree to which they alter the targeted psychological trait. Consider the earlier case of Bryony. We supposed that her sympathy-enhancing intervention increased her moral conformity because it enabled her to better conform to her moral reasons to provide humanitarian aid. However, it is easy to imagine that, had her sympathies become a little stronger, she might instead have conformed less well. Perhaps she would then have been overwhelmed, and thus paralysed, by those sympathies. By contrast, when an individual increases her moral conformity by augmenting her moral-epistemic resources, her moral conformity will generally be more robust across different magnitudes of change. For example, it will normally be the case that her moral conformity would also have increased had she augmented her moral-epistemic resources to a slightly greater or slightly lesser degree.

There are, then, at least two *prima facie* reasons to suppose that brute conformity enhancements will produce less reliable moral conformity than typical deliberative alternatives: it may be that the moral conformity induced by brute conformity enhancements is both more contingent on the circumstances and more sensitive to the magnitude of the transformation that the enhancement induces.[32] If the moral worth of an act that conforms to morality is a function of the reliability of that conformity, these considerations will help to explain why brute conformity enhancements are less conducive to moral worth than their deliberative counterparts.

In response to this attempt to justify the Moral Worth Claim, an Aristotelian thought concerning the connection between moral conformity and moral knowledge might be inserted into the discussion. On one Aristotelian account of moral education, moral knowledge is acquired in part *through* conforming to morality. In the moral sphere, we 'learn by doing'.[33] If this account is correct, we should expect that brute conformity enhancements, like paradigmatic deliberative ones, will typically produce moral knowledge, and thus, other things being equal, augment our moral-epistemic resources. Brute conformity enhancements by definition enhance moral conformity, and moral conformity is itself conducive to the acquisition of moral knowledge. Thus, it might seem, the problems about unreliability have been overstated. Moral knowledge reliably produces moral conformity, and brute conformity enhancements tend to produce moral knowledge.

[32] Note that the point here is not that those who have undergone brute conformity enhancements fail to act on reliable *motives*. The problem, rather, is that, in someone who acts on a reliable motive as a result of having undergone some brute intervention, the presence of the reliable motive will itself be problematically contingent on circumstances.

[33] See, for the classic presentation of this interpretation, Burnyeat [34].

This reply engages in a problematic form of bootstrapping, however. We were interested in assessing the reliability of the moral conformity produced by brute conformity enhancements. It will not help, in addressing this issue, to maintain that *insofar as these interventions do increase moral conformity* they will also produce moral knowledge, which is a reliable promoter of moral conformity. Nothing has been done to resolve the initial unreliability: the unreliability with which the brute psychological change induced by the intervention produces moral conformity.

There is, however, an alternative, more persuasive response to the present attempt to justify the Moral Worth Claim. The response begins with the thought that brute conformity enhancements could operate by lowering barriers to moral knowledge, moral understanding or moral judgment. For example, they could improve moral conformity by attenuating some emotion or desire that acts as a barrier to clear thinking or vivid imagination, both of which plausibly facilitate the acquisition of moral knowledge, understanding and judgment. Of course, even brute enhancements of this sort would, in an important sense, be unreliable. Though they might in fact augment one's moral-epistemic resources, they would not reliably do so. For example, though pharmacologically augmenting the capacity for vivid imagination, by lowering some barrier to it, might sometimes produce better moral judgment, there are circumstances in which it would fail to do so. Consider a case in which, whatever an agent will do, the consequences will be horrific. The agent's task is to select which of two serious atrocities to prevent, say. In this sort of case, one might think that vividly imagining the outcomes would serve only to traumatise the agent in a way that is likely to cloud the agent's moral judgment.

Note, however, that typical *deliberative* conformity enhancements are unreliable in precisely the same way. Consider an agent, Dennis, who has racist beliefs, but on some level recognises that his beliefs are racist and therefore objectionable. Suppose Dennis seeks to confront his racism through deliberation—say, by reading and reflecting on the *Adventures of Huckleberry Finn*. Such deliberation might well increase his moral knowledge, and might thereby improve his moral conformity. However, even if we accept that moral knowledge reliably produces moral conformity, there is a sense in which Dennis' moral conformity, following his deliberation, is accidental. It is accidental in that his reading and reflecting on *Huckleberry Finn* was not *guaranteed* to produce moral knowledge, at least if Dennis is an ordinary person. *Perfect* moral deliberation might invariably produce moral knowledge. But no ordinary person is a perfect moral deliberator—someone who, when he engages in moral deliberation, always does so perfectly. Moreover, we have particular reasons to doubt Dennis' deliberative abilities: by hypothesis, he has racist beliefs, and we might worry that these beliefs, or emotional reactions that may underpin them, will infect his deliberation. For example, perhaps they will lead him, when deliberating, to selectively read the evidence in a way that helps to maintain those beliefs.

More generally, an agent's moral deliberation will enhance that agent's moral-epistemic resources—her moral knowledge, understanding and judgment—only when it goes well, and this is something that cannot, in an ordinary person, be relied upon. It depends, for example, on the absence of certain external impediments to good deliberation (the absence of temptations and distractions, say). So even where deliberation *does* augment an agent's moral-epistemic resources, and thus leads to greater moral conformity, there is an important sense in which that moral conformity is not reliable. The acquisition of the moral-epistemic resources was itself contingent on favourable circumstances for deliberation.

This is the same problem that we identified in the case of *brute* conformity enhancements. Though such enhancements may augment the moral-epistemic resources, which in turn reliably produce moral conformity, they do not reliably do so. Their doing so is contingent on favourable circumstances obtaining.

There may, of course, be some individuals—those particularly disposed to sound moral deliberation—in whom moral deliberation will produce more reliable moral conformity than any alternative brute conformity enhancement and will thus, perhaps, be the route to moral conformity most conducive to moral worth. But this does not help the proponent of the Moral Worth Claim, who maintains that *typical* deliberative conformity enhancements are more conducive to moral worth than any brute alternatives that might plausibly be developed in the medium term future. Given that the moral conformity produced by both kinds of conformity enhancement is unreliable in the same sort of way, there seems little reason to suppose that typical

deliberative conformity enhancements will produce more reliable moral conformity than brute conformity enhancements of this sort.

Conclusions

John Harris holds that we have reason to adopt typical deliberative conformity enhancements in preference to brute conformity enhancements. I granted that the Moral Worth Claim supports this view and considered three attempts to justify that Claim. Each of these attempts drew on ideas that have often been associated with a Kantian approach to moral worth. The first attempt appealed to the standard Kantian view that to have moral worth, the action must be done from the motive of duty. I argued that this attempt fails because, on an othodox understanding of acting on the motive of duty, brute conformity enhancements could bring it about that one acts on precisely this motive. The second attempt appealed to the view that moral effort confers moral worth along with the view that brute conformity enhancements would be less effortful than typical deliberative alternatives. I argued that, if this attempt succeeds in establishing the Moral Worth Claim, it does so only on the assumption that, leaving aside considerations of moral worth, we have more reason to pursue deliberative conformity enhancements than brute alternatives. However, this assumption deprives the Moral Worth Claim of its interest as a way of defending Harris view. Finally, the third attempt appealed to the view that moral worth requires reliable moral conformity along with the claim that brute conformity enhancements would produce less reliable moral conformity than typical deliberative alternatives. I responded to this attempt by arguing that both kinds of conformity enhancement are unreliable in the same sort of way.

Where does this leave Harris' view that we should prefer deliberative conformity enhancements to brute alternatives? Plainly this depends on (i) whether it is possible to establish the Moral Worth Claim via some route that I have not pursued here, and (ii) the persuasiveness of Harris' other concerns about brute conformity enhancements.[34] However, the failure of the three attempted justifications for the Moral Worth Claim considered here should, I think, at least significantly lower our credence in Harris' view. This is because I believe that those three attempts are perhaps the most promising means of establishing Harris' view.

[34] For responses to these concerns, see DeGrazia [13], Douglas [9, 10], Savulescu and Persson [35] and Savulescu, Douglas and Persson, (2013), Autonomy and the ethics of biological behaviour modification, unpublished.

Acknowledgments I thank Kelly Sorensen, Ingmar Persson, Peter Railton, three anonymous reviewers, and audiences in Belgrade and Stockholm for comments on earlier versions of this paper. I also thank the Wellcome Trust [grant number WT087211] for their funding, and John Harris and Robert Sparrow for sharing unpublished work.

References

1. Lifton, Robert Jay. 1986. *The Nazi doctors: Medical killing and the psychology of genocide*. New York: Basic Books.
2. Darley, J.M. 1992. Social organization for the production of evil. *Psychological Inquiry* 3(2): 199–218.
3. Glover, Jonathan. 1999. *Humanity: A moral history of the Twentieth Century*. London: Jonathan Cape.
4. Harris, John. 2011. Moral enhancement and freedom. *Bioethics* 25(2): 102–111.
5. Harris, John. 2012. What it's like to be good. *Cambridge Quarterly of Healthcare Ethics*. doi:10.1017/S0963180111000867.
6. Harris, John. 2012. 'Ethics is for bad guys!' Putting the 'moral' into moral enhancement. *Bioethics*. doi:10.1111/j.1467-8519.2011.01946.x.
7. Persson, Ingmar, and Julian Savulescu. 2008. The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of Applied Philosophy* 25(3): 162–177.
8. Persson, Ingmar, and Julian Savulescu. 2010. Moral transhumanism. *The Journal of Medicine and Philosophy* 35(6): 656–669.
9. Douglas, Thomas. 2008. Moral enhancement. *Journal of Applied Philosophy* 25(3): 228–245.
10. Douglas, Thomas. 2011. Moral enhancement via direct emotion modulation: A reply to John Harris. *Bioethics*. doi:10.1111/j.1467-8519.2011.01919.x.
11. Faust, Halley S. 2008. Should we select for genetic moral enhancement? A thought experiment using the MoralKinder (MK+) haplotype. *Theoretical Medicine and Bioethics* 29(6): 397–416.
12. Crockett, Molly J., Luke Clark, Marc D. Hauser, and Trevor W. Robbins. 2010. Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences* 107(40): 17433–17438.

13. DeGrazia, David. 2013. Moral enhancement, freedom, and what we (should) value in moral behavior. *Journal of Medical Ethics*.
14. von Leibniz, Gottfired Wilhelm. 1973 [1695]. New system, and explanation of the new system. In *Philosophical writings*, ed. G. H. R. Parkinson, trans. Mary Morris. London: Dent.
15. Cohen Kadosh, Roi, Sonja Soskic, Teresa Iuculano, Ryota Kanai, and Vincent Walsh. 2010. Modulating neuronal activity produces specific and long-lasting changes in numerical competence. *Current Biology* 20(22): 2016–2020.
16. Karim, Ahmed A., Markus Schneider, Martin Lotze, Ralf Veit, Paul Sauseng, Christoph Braun, and Niels Birbaumer. 2010. The truth about lying: Inhibition of the anterior prefrontal cortex improves deceptive behavior. *Cerebral Cortex* 20(1): 205–213.
17. Hursthouse, Rosalind. 1991. Virtue theory and abortion. *Philosophy & Public Affairs* 20(3): 223–246.
18. Jotterand, Fabrice. 2011. 'Virtue engineering' and moral agency: Will post-humans still need the virtues? *American Journal of Bioethics—Neuroscience* 2(4): 3–9.
19. Sorensen, Kelly. 2004. The paradox of moral worth. *Journal of Philosophy* 101(9): 465–483.
20. Harris, John, and Sarah Chan. 2010. Moral behavior is not what it seems. *Proceedings of the National Academy of Sciences* 107(50): E183.
21. Johnson, Robert N. 1996. Kant's conception of merit. *Pacific Philosophical Quarterly* 77: 313–337.
22. Arpaly, Nomy. 2003. *Unprincipled virtue: An inquiry into moral agency*. New York: Oxford University Press.
23. Henson, Richard G. 1979. What Kant might have said: Moral worth and the overdetermination of dutiful action. *Philosophical Review* 88(1): 39–54.
24. Wood, Allen. 1999. *Kant's ethical thought*. Cambridge: Cambridge University Press.
25. Audi, Robert. 2009. Moral virtue and reasons for action. *Philosophical Issues* 19(1): 1–20.
26. Hurka, Thomas. 2001. *Virtue, vice, and value*. New York: Oxford University Press.
27. Kant, Immanuel. 1964 [1785]. *Groundwork of the metaphysic of morals*. trans. H. J. Paton, 1st Harper torchbook ed. New York: Harper & Row.
28. Herman, Barbara. 1981. On the value of acting from the motive of duty. *Philosophical Review* 90(3): 359–382.
29. Herman, Barbara. 1996. Making room for character. In *Aristotle, Kant, and the Stoics: Rethinking happiness and duty*, ed. Stephen P. Engstrom and Whiting Jennifer. Cambridge: Cambridge University Press.
30. Sorensen, Kelly. 2010. Effort and moral worth. *Ethical Theory and Moral Practice* 13(1): 89–109.
31. Kant, Immanuel. 1991 [1797]. *The metaphysics of morals*, trans. M. Gregor. New York: Cambridge University Press.
32. Markovits, Julia. 2010. Acting for the right reasons. *Philosophical Review* 119(2): 201–242.
33. Hills, Alison. 2009. Moral testimony and moral epistemology. *Ethics* 120(1): 94–127.
34. Burnyeat, Miles F. 1980. Aristotle on learning to be good. In *Essays on Aristotle's ethics*, ed. A.O. Rorty, 69–92. Berkeley: University of California Press.
35. Savulescu, Julian, and Ingmar Persson. 2012. Moral enhancement, freedom, and the god machine. *Monist* 95(3): 399–421.