IMPLANTED DESIRES, SELF-FORMATION AND BLAME

BY MATTHEW TALBERT

JOURNAL OF ETHICS & SOCIAL PHILOSOPHY

VOL. 3, NO. 2 | AUGUST 2009
URL: WWW.JESP.ORG

Implanted Desires, Self-Formation and Blame

Matthew Talbert

SOMETIMES A PERSON SEEMS less blameworthy for her actions when we learn how she came to be the sort of person she is. Consider, for example, how often novelists and scriptwriters invite us to reconsider our feelings toward an unlikable character by showing us crucial facts about the formative influences to which that character was subject. In the philosophical literature on responsibility, one of the best-known accounts of this phenomenon is Gary Watson's discussion of the convicted murderer Robert Alton Harris.[1] Harris committed brutal crimes, but our initial reactions toward him are called into question when we learn about the abuses he suffered as a child.

We see Harris initially as a victimizer, but a more complete account of his story reveals him to have also been a victim. Placing Harris simultaneously in both these categories makes it difficult to sustain unequivocal emotional responses toward him. Other factors evoke a similar reconsideration of our attitudes toward wrongdoers. In some cases, we may imagine that had *we* been exposed to certain influences, we would have turned out like the person we condemn. This awareness may make our condemnation seem inappropriate because the bad behavior in question now seems partly a function of bad moral luck. Alternatively, we may suspect that some formative situations distort a person's understanding of right and wrong. This happens in Susan Wolf's "JoJo" example.[2] JoJo was raised by a vicious dictator and as a consequence he lacks the resources to recognize the status of his own immoral actions. To the extent that moral understanding is a condition on responsibility, JoJo is an unfit target for blame.

There are, then, different ways in which considerations about an agent's past can bear on his present blameworthiness. Note, for instance, that in Wolf's example the facts about JoJo's past are only indirectly related to his exemption from blame. What really matters for Wolf is JoJo's incapacity in the face of moral reasons. In this paper, I will argue against a picture of responsibility that is somewhat different from Wolf's – a picture on which an agent's past has a *direct* impact on his blameworthiness, unmediated by considerations about moral understanding. I will call the view I have in mind "historicism." Historicism contends, roughly, that if an agent is not responsible for the fact that she has certain action-guiding values and desires, then she is not fully responsible for acting on those desires and values. Historicist theories may differ in their details, but a central implication of the view is that a given agent might be morally responsible for her behavior while that

[1] Gary Watson, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in G. Watson, *Agency and Answerability: Selected Essays* (Oxford: Clarendon, 2004), pp. 219-259.
[2] Susan Wolf, "Sanity and the Metaphysics of Responsibility," in F. Schoeman, *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (Cambridge: Cambridge University Press, 1987), pp. 46-62.

agent's psychological twin is not responsible for apparently identical behavior. According to the historicist, this difference in responsibility may have nothing to with any difference in how the two agents produced their actions; instead, the difference in responsibility may be due simply to the fact that the first agent, but not the second, fulfilled whatever historical conditions properly apply to moral responsibility.[3]

In the philosophical literature on the relationship between personal history and moral responsibility, realistic cases of childhood abuse or corrupting social contexts often give way to fanciful scenarios involving Skinnerian conditioning and other overt manipulations of subjects' desires and values. Historicists offer these extreme cases as instances in which an agent's history seems obviously relevant to his responsibility. One well-known example of this sort, devised by Alfred Mele, involves a woman named Beth who is subjected to covert psychological manipulation that turns her into a psychological duplicate of Charles Manson.[4] By hypothesis, Manson is morally responsible for his value-guided, vicious deeds. The question is whether Beth is also responsible for acting on her strangely acquired values.

Against the historicist, I argue below that it would be reasonable to hold someone like Beth morally responsible for her actions. I claim, in particular, that if Beth were to maliciously injure another person after her manipulation, it would be appropriate to respond to Beth with the emotions, attitudes and demands that characterize moral blame.[5] Beth could be blameworthy for some of her post-manipulation actions, I argue, because the origin of her values and dispositions does not entail that her actions fail to display the core features of blameworthy behavior. The general conception of blame I employ below is inspired by P. F. Strawson's influential interpretation of the negative reactive emotions that characterize blame as "reactions to the quality of others' wills towards us, as manifested in their behaviour: to their good or ill will or indifference or lack of concern."[6] Since, as I shall argue, Beth's ma-

[3] As I suggest in the text, some historical conditions require that a responsible agent have made certain contributions to his own development in order to be responsible for his present actions; other historical conditions stipulate that certain responsibility-undermining influences must have been *absent* from a responsible agent's development. David Zimmerman discusses the difference between these "negative" and "positive" historical conditions in detail. See Zimmerman, "That Was Then, This is Now: Personal History vs. Psychological Structure in Compatibilist Theories of Autonomy," *Noûs* 37 (2003), pp. 646-48.

[4] Alfred Mele, *Autonomous Agents: From Self-Control to Autonomy* (New York: Oxford University Press, 1995), pp. 156-62. I rely on Mele's provocative examples and thought-experiments in several places, but in this essay I am mainly concerned to offer a counterpoint to general historicist intuitions rather than to fully engage Mele's specific (and detailed) articulation of historicism.

[5] There is more to moral responsibility than blameworthiness, but if a person is an apt target for the responses that constitute blame, then she is a morally responsible agent in at least one significant sense.

[6] P. F. Strawson, "Freedom and Resentment," in G. Watson, *Free Will*, Second Edition (New York: Oxford University Press, 2003), pp. 72-93. It is quite possible that Strawson would not agree with my application of his view. Strawson suggests that we should suspend blame in

nipulation does not render her incapable of possessing ill will toward others and guiding her behavior on that basis, she is capable of committing actions to which moral blame is a reasonable response. This general Strawsonian conception of blame is familiar and attractive to many compatibilists about moral responsibility; however, not all compatibilists accept the ahistorical approach I advocate. In the last section of this paper, I argue that compatibilists have particular reason to eschew historical conditions on responsibility.

I will return to manipulation cases like Beth's shortly, but it will be useful first to consider Harry Frankfurt's compatibilist account of moral responsibility. Frankfurt's position is interesting both because it has been criticized by historicists and because it points to a relation between desires and reasons that will be useful in analyzing the manipulation examples that supposedly favor historicism.

I. Identification and Personal History

First, consider Frankfurt's example of a *willing* drug addict. Given the severity of this addict's addiction, he will take his drug regardless of the higher-order attitudes he forms about the desires that characterize his addiction. The willing addict therefore lacks freedom of the will in Frankfurt's sense: he will be moved by his desire to take his drug regardless of whether he wants to be so moved.[7] However, the willing addict is *willing*: he identifies with his addictive desire and affirms it from a higher-order perspective. Thus, even though the willing addict lacks an important sort of freedom, we may feel that he is still blameworthy for his drug use to the degree that he has made his will his own by reflectively endorsing the desires that move him.

By contrast, a similarly addicted agent might take a stand against, and find herself alienated from, her addictive desire and so may fail to be fully responsible for an action motivated by that desire.[8] If responsibility fails to obtain in the case of the unwilling addict, it is because of the higher-order stance this second addict takes with respect to her addiction. On Frankfurt's view, then, responsibility is crucially related to whether an agent identifies herself with, or withdraws herself from, the desires that move her.

A number of authors have criticized the role that higher-order dispositions play in Frankfurt's account of responsibility, and these criticisms have often had a historicist cast. For example, Frankfurt's view is the pre-eminent

---

the case of one whose "mind has been systematically perverted" (p. 78) or who has been "peculiarly unfortunate in his formative circumstances" (p. 79), and he may have felt the same about someone who has been manipulated in the way Beth has.

[7] Harry Frankfurt, "Freedom of the Will and the Concept of a Person," in H. Frankfurt, *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988), pp. 23-5.

[8] Frankfurt says that in some cases a desire "moves [the agent] to act against his own will . . . In this respect it is alien to him, which may justify regarding him as having been moved passively to do what he did by a force for which he cannot be held morally responsible." Frankfurt, "Three Concepts of Free Action," in *The Importance of What We Care About*, p. 48.

instance of what John Martin Fischer and Mark Ravizza call "mesh theories" of responsibility. On a mesh theory, an agent is morally responsible for an action as long as there is an appropriate mesh between how the agent wants to be moved and the desires that actually produce her actions.[9] Thus, mesh theories are ahistorical: responsibility is simply a matter of the contemporary structure of an agent's will and it does not matter how this structure came about. Fischer and Ravizza argue that mesh theories (and Frankfurt's account in particular) are implausible just because these theories apparently allow that an agent's higher-order endorsements of his desires might result from *any* process and still contribute to the agent's moral responsibility. For Fischer and Ravizza, such accounts of responsibility are incomplete because they do not discriminate against intuitively "responsibility-undermining" processes (such as brainwashing) by which an agent's will might acquire its overall structure.[10]

Frankfurt has not been moved by this type of criticism and continues to insist that even an agent whose psychology has been manipulated by maximally invasive procedures is responsible for her actions as long as she has the right higher-order dispositions toward the desires that move her.[11] I agree with the substance of Frankfurt's conclusion, but given the resistance that Frankfurt has encountered on this subject, a shift in emphasis may aid the non-historicist in the presentation of her case. More progress may be made in motivating an ahistorical outlook if we focus not on how an agent is disposed toward his desires but on whether a given desire helps to explain the agent's behavior because he took the desire as a reason for action.[12]

One consideration in favor of this shift in focus is that presenting the issue, as Frankfurt does, in terms of higher-order desires naturally invites a standard historicist critique. On Frankfurt's account, part of the significance of an agent's capacity to identify (or not) with her desires is that this allows her to define herself for the purposes of moral assessment. By identifying with a desire, an agent makes that desire her own so that when she acts on

[9] John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1998), p. 184.
[10] Fischer and Ravizza, *Responsibility and Control*, p. 196.
[11] Harry Frankfurt, "Reply to John Martin Fischer," in S. Buss and L. Overton, *Contours of Agency: Essays on Themes from Harry Frankfurt* (Cambridge, MA: MIT Press, 2002), pp. 27-31.
[12] This approach is inspired by Michael Bratman's analysis of the concept of identification in terms of a "decision," on the part of an agent, to treat a desire as a reason to act. See Bratman, "Identification, Desire, and Treating as a Reason," in M. Bratman, *Faces of Intention: Selected Essays on Intention and Agency* (Cambridge: Cambridge University Press, 1999), pp. 185-206. The shift in focus I advocate does not necessarily conflict with Frankfurt's hierarchical model of responsibility; after all, taking a desire as a reason may itself involve a higher-order perspective on the part of an agent. This is consistent with my aim, which is not to produce an account that is substantively opposed to Frankfurt's, but rather to encourage a way of accessing the debate about the relevance of personal history to moral responsibility that avoids (or at least forestalls) certain modes of resistance that Frankfurt has regularly encountered.

the desire, she shows where she stands, morally speaking. However, we might worry that identification can play this role only if an agent's higher-order preferences, through which identification is accomplished, already belong to the agent in an authentic way. Thus, it may seem natural to ask – as Frankfurt's historicist critics do – where the psychological states involved in identification got their authority to speak for the agent. But, at this point, the historicist will insist – as Fischer and Ravizza do – that certain stories about how an agent came by her higher-order preferences actually inhibit the agent's capacity to authoritatively identify with her lower-order desires.[13]

Frankfurt's picture of ascending orders of reflection makes the historicist's concern a natural one because it seems that every level of reflection must receive its authority to speak for the agent from a still-higher order of reflection.[14] Thus, as far as Frankfurt's theory goes, it always seems an open question where a higher-order desire got its authority to play the role it plays in an agent's psychic economy. However, if we shift our focus to the question of whether an agent takes a desire as a reason to act a certain way, then there is less room for the historicist's question to arise. We care about the contexts in which a person takes his desires to be reasons in part because we care how agents value other things in comparison with the satisfaction of their desires. I care, for example, whether you take yourself to have good reason to satisfy a certain desire when you know that doing so will cause me an unjustified injury. How you settle this conflict between your desires and my well-being tells me something about how you are disposed toward me and about whether you bear good or ill will toward me; and, on the view employed here, it is judgments about the quality of will you show me that are relevant to blame. But these judgments can be accurate, and support the emotions involved in blame, without me knowing whether you played a role in making yourself the sort of person who weighs reasons as you do.[15]

---

[13] For another criticism of Frankfurt along these lines, see John Christman, "Autonomy and Personal History," *Canadian Journal of Philosophy* 21, (1991): pp. 1-24.

[14] Frankfurt has attempted in various places to avoid this regress. He says, for example, that when "a person identifies himself *decisively* with one of his first-order desires, this commitment 'resounds' throughout the potentially endless array of higher orders." Frankfurt, "Freedom of the Will," p. 21. I will assume, however, that from the historicist's perspective, Frankfurt is still saddled with a regress problem.

[15] Pamela Hieronymi makes a similar point when she says that the emotional reactions that characterize blame are "sensitive to just those facts that make the wrongdoing interpersonally important . . . . They are not sensitive to facts that do not change, in some understandable way, the significance of the wrongdoing for one's interpersonal relations." Hieronymi, "The Force and Fairness of Blame," *Philosophical Perspectives* 18, (2004): p. 135. And, as Hieronymi points out, facts about an agent's history (and other related considerations) need not alter the significance of a person's actions for us because these facts need not alter our judgments about the quality of the agent's will. Of course, the historicist might agree that a manipulated agent can express ill will through her actions but deny that the will in question is properly attributed to the agent. That is, the historicist may be concerned that the entire evaluative perspective, from which a manipulated agent's actions issue, is not really the agent's own. I

Another reason for moving away from Frankfurt's emphasis on the hierarchical nature of the will is that this emphasis can distract us from the fact that agents are often blameworthy even when they disapprove of the first-order impulses that move them. There are cases, no doubt, where an agent is so alienated from his desires that we ought not to hold him responsible for his actions. Such an agent might be, as Frankfurt puts it, "helplessly violated by his own desires."[16] But while this strong conception of alienation may apply to instances of mind control or compulsive behavior, these are rare and seriously aberrant cases. Most cases in which an agent is conflicted about, or withdrawn from, her motivations are much less serious and do not necessarily call responsibility into question.

In a recent presentation of his views, Frankfurt says that a rejected, "externalized impulse or desire may succeed, by its sheer power, in defeating us and forcing its way."[17] As noted above, when a person is overpowered in this way, her responsibility is plausibly called into question. But Frankfurt's subsequent example seems unlikely to be an instance of alienation in this strong sense. It might be the case, says Frankfurt, that "[i]nstead of being moved by the warm and generous feelings that he would prefer to express, a person's conduct may be driven by a harsh envy, of which he disapproves but that he has been unable to prevent from gaining control."[18]

In a case like this, we might accept a person's claim that she disapproves of the envy she feels, but if she is moved by this envy, then we need not accept the additional claim that she has been overwhelmed by an external power. It is not difficult to imagine a scenario like the one Frankfurt describes in which the role the agent's desires and feelings play in her action seem to amount to the agent's participation in that action, even though we also believe that she disapproves of these features of her psychology.

Suppose, for instance, that I have some obligation to you, but that I cannot help being distracted from this obligation by my desire to engage in some enjoyable, but trivial, pastime.[19] Now I may decide that what I ought to do under the circumstances is to attend to my obligation, and I may disapprove of the fact that the pastime persistently presents itself to me as an appealing alternative. In the end, I may fail to meet my obligation and I may truthfully report that my reason for so acting was the enticement of the pastime. But it would be odd, I think, to say in this case that I am absolved of blame just because I have acted contrary to my own commitments about what reasons I have.

---

will address this sort of worry in Section IV when I discuss the problem of cross-manipulation identity.

[16] Frankfurt, "Freedom of the Will," p. 17.

[17] Harry Frankfurt, *Taking Ourselves Seriously & Getting It Right* (Stanford, CA: Stanford University Press, 2006), p. 14.

[18] Frankfurt, *Taking Ourselves Seriously*, p. 15.

[19] This example is modeled on one from T. M. Scanlon, *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press, 1998). p. 36.

I may well be blameworthy here because my action reveals an insufficient responsiveness to my obligations. My considered judgment about what I have reason to do shows that I am partially sensitive to my obligations, but this was apparently insufficient to keep me from being distracted, and ultimately motivated, by a desire that pulls me in an opposing direction. I may disapprove of what I have done and regret that I cannot help seeing my desire to engage in a trivial pastime as a reason to do so, but this does not mean that I have been overwhelmed by the enticing distraction in a way that calls my control over my behavior, or my responsibility, into question.

Gary Watson has suggested that when we succumb to a powerful desire – even one to which we are officially opposed – we are typically "not so much overpowered by brute force as seduced."[20] To be effectively seduced is to be converted – to some extent – to a different point of view; thus, seduction is accomplished, at least partly, through the participation of the one who is seduced. An agent moved by the seductive power of a desire may remain conflicted about what he is doing, yet because of the role the desire plays in explaining his action, he is not properly described as alienated from his action in a way that undermines responsibility. This is why "I was seduced" may count as an explanation of behavior, but it is not typically a compelling excuse.

II. Reasons and Desires

So far, I have suggested that when we give an account of an agent's responsibility for acting on a desire, we should focus on whether the desire played the role of a reason for the agent. This is because people's judgments about reasons tell us about their normative commitments and their esteem for others, and these facts play a crucial role in determining whether a person is open to the negative emotional responses that characterize moral blame. However, my comments about reasons and desires so far have been insufficiently nuanced.

To say that we care whether an agent takes a desire as a reason assumes that desires can be reasons and perhaps implies that desires are a primary source of motivation. However, these assumptions can be questioned. T. M. Scanlon and Warren Quinn have both argued that taking oneself to have a reason – rather than possessing a desire – is basic to motivation, and that desires themselves are best interpreted in terms of the phenomenon of seeing something as a reason.[21] If we find this proposal appealing, then instead of saying that an agent decides that a given desire is a reason to φ, we should say

[20] Gary Watson, "Disordered Appetites: Addiction, Compulsion, and Dependence," in *Agency and Answerability*, p. 71.
[21] Scanlon, *What We Owe to Each Other*, pp. 37-55; Warren Quinn, "Putting Rationality in Its Place," in W. Quinn, *Morality and Action* (Cambridge: Cambridge University Press, 1993), pp. 228-255.

that an agent has a desire to φ, properly speaking, if, among other things, it seems to the agent that there are considerations in favor of φ-ing.

Of course, a person can be motivated to φ even though she does not see anything in favor of φ-ing. Consider an example of Quinn's that illustrates this point. Quinn imagines himself in a "strange functional state that disposes [him] to turn on radios that [he sees] to be turned off."[22] What makes this state strange is that a given radio being turned on does not achieve anything Quinn values: he is not interested in listening to music or news, or even in simply avoiding silence. While we might refer to this motivational state in order to explain why he turned on a radio, Quinn claims that merely being in this state does not give him "even a *prima facie* reason to turn on radios" and "does not make the act sensible, except insofar as resisting the attendant disposition is painful."[23] It is perhaps not quite appropriate to call Quinn's odd motivational state a *desire*, so we can refer to this sort of non-rationalizing motivational state as an *urge*.[24]

The distinction between urges and desires is applicable to Frankfurt's addiction cases (as well as to the manipulation cases I consider in the next section). If an addict acts on a non-rationalizing urge, then blame may be inappropriate because the agent's action does not indicate anything relevant about her judgments about what counts as a reason.[25] Perhaps this is what happens in the case of the unwilling addict, who is not responsible for acting on her addiction. What explains the absence of responsibility in this case is not simply the fact that the unwilling addict would rid herself of her urge if she could, but that her action does not issue from morally relevant judgments about reasons. If, however, the unwilling addict acts on a rationalizing desire – if she is enticed by considerations that seem to her to count in favor of taking the drug – then she may be blameworthy even though her being enticed in this way conflicts with her considered opinion about what reasons she has.[26]

---

[22] Quinn, "Putting Rationality in Its Place," p. 236.

[23] Quinn, "Putting Rationality in Its Place," pp. 236-37.

[24] Scanlon says that the aspect of "seeing something as in some way worth doing, or worth bringing about, is what differentiates desires from mere urges." Scanlon, "Reasons, Responsibility, and Reliance: Replies to Wallace, Dworkin, and Deigh," *Ethics* 112 (2002), p. 508. Gary Watson suggested a similar distinction when he considered a "woman who has a sudden urge to drown her bawling child in the bath," but of whom it is false to say that she "values her child's being drowned." Watson, "Free Agency," in *Agency and Answerability*, p. 19.

[25] However, acting on a non-rationalizing urge will not always exempt an agent from blame. In some such cases – depending on what the agent is motivated to do – blame may be undermined only if the urge in question is irresistible, but the fact that an agent performs an action on the basis of an urge does not mean that he could not have resisted that urge.

[26] Of course, an unwilling addict might act on a reason that stems from her interest in avoiding the pain involved in not taking her drug. Sarah Buss makes this point in "Autonomy Reconsidered," in P. French, T. Uehling, Jr., H. Wettstein, *Midwest Studies in Philosophy* 19 (Notre Dame, Ind.: University of Notre Dame Press, 1994), pp. 95-121. This may make an addict's actions understandable in a way that limits blameworthiness. Thus, the fact that an

III. Manipulation Cases

We are now in a position to assess some of the thought experiments meant to elicit historicist intuitions. In these cases, a person's desires or values are directly implanted in her instead of being acquired as part of a process in which the agent participates. I will argue that, as far as responsibility is concerned, what matters in these cases is not how an agent came to have her desires but whether, when she acts to satisfy a desire, she acts because of considerations that she counts in favor of so acting. On this analysis, manipulation is typically irrelevant to an assessment of moral blameworthiness.

However, it must be admitted that the victims in *some* manipulation scenarios should not be held responsible for their actions. Consider what happens in the Cold War thriller, *The Manchurian Candidate*. There, Chinese scientists subject U. S. military personnel to brainwashing techniques that achieve certain ends – e.g., the production of verbalizations and bodily movements – by circumventing the subjects' reflective capacities, desires and values. Clearly, responsibility is also bypassed in these cases. But this is because the manipulated agents' capacities for reflective self-government do not play a role in the production of their actions; thus, in *Manchurian Candidate*-type cases, an agent's actions do not express the sort of normative commitments to which blame properly responds. While responsibility *is* undermined here, this does not support a historicist conclusion because the absence of responsibility does not follow from the victim's desires being manipulated or implanted. In fact, in *Manchurian Candidate*-type manipulations, an agent's desires are not manipulated at all and they do not play their customary role in the formation of intention and the production of action.

A more compelling manipulation case, from the historicist's point of view, would be one in which an agent's desires or values are themselves manipulated. Imagine, for example, that an uncharacteristic desire to commit a crime is directly implanted into a subject with the aim that she should act on that desire. Such an example is, no doubt, technologically implausible, but it is more important here to note the psychological implausibility of supposing that such a manipulation could bring about its aim. If what is implanted is a single desire that is not accompanied by altered tendencies to see considerations as counting in favor of acting on that desire, then it is unclear that the manipulation could bring about an action that the manipulated agent would otherwise have been disinclined to perform.

And even if we imagine that a lone, dissonant motivation could be implanted in an agent such that it is overwhelmingly strong and capable of causing action, this would still not necessarily support a historicist conclusion. While such a case would be another instance of responsibility-undermining

agent acts for some reason or other is not enough to ground moral blame; the content of her reasons also matters.

manipulation, responsibility would be undermined here only because the overwhelmingly strong – but non-rationalizing – implanted desire amounts to the sort of "urge" introduced above. Such a case would be similar to the instances of extreme (and responsibility-undermining) alienation to which Frankfurt draws our attention. Again, this would not be a case of an agent who is made blameless because her otherwise normal action issues from a manipulation. Rather, and regardless of whether it issues from manipulation, the action in this case bears little resemblance to normal, responsible agency. Thus, this case would not support historicism either.

In the two previous cases, a non-historicist can allow that the manipulated agents are not responsible without conceding historicism's larger point that manipulated agents are not responsible *just because* their values have been forcibly, and unshakably, implanted in them. This is because in both these scenarios responsibility is undermined for reasons that are conceptually independent of the fact of manipulation. What the historicist needs to offer is a manipulation example that features normal, value-guided action of the sort for which agents are typically held responsible. I turn now to consider such cases: Al Mele's thought experiments about "Beth."

Beth is a less than ideally productive philosopher and her dean would like Beth to be more like Ann, who is a very productive philosopher. The dean arranges for a team of brainwashers to secretly manipulate Beth so that she becomes psychologically identical (in the relevant respects) to Ann. After the brainwashing, Beth has Ann's "hierarchy of values" and she satisfies the structural requirements that Frankfurt puts on responsibility: when she "reflects on her preferences and values, Beth finds that they fully support a life dedicated to philosophical work, and she wholeheartedly embraces such a life and the collection of values that supports it."[27]

According to Mele:

> The salient difference between Ann and Beth is that Ann's practically unsheddable values were acquired under her own steam, whereas Beth's were imposed upon her. Ann autonomously developed her values . . . Beth plainly did not. . . . Ann and Beth make equal use of the relevant, unsheddable values in 'governing' their mental lives. But in Beth's case, one is inclined to view this as *ersatz* self-government. The dean and his cronies seized control of the direction that her life would take . . . . Behind the facade of self-government, external governors lurk . . . [28]

The historicist will conclude that since Beth's self-government is only apparent, her responsibility is only apparent as well.[29]

---

[27] Mele, *Autonomous Agents*, p. 145.

[28] Mele, *Autonomous Agents*, pp. 155-56.

[29] As the quoted passage indicates, Mele's explicit concern is with autonomy, which is separable from concerns about responsibility. However, in the larger context of Mele's book, it is clear that he thinks that Beth's responsibility, as well as her autonomy, is called into question by her manipulation. This is also clear in Mele's recent treatment of the Beth/Ann case, as

IV. Manipulation and Self-Government

Extreme examples like the one involving Beth are designed to make a historicist conclusion seem inevitable. Less exotic cases, involving more prosaic methods of introducing preferences and values into an agent's psychology, might not lead so obviously to a historicist conclusion. However, even in the extreme case, there is reason to respond to a brainwashed Beth as we would respond to a non-brainwashed Ann, praising or blaming her for the effects of a zealous commitment to philosophical work.

One reason to draw this conclusion has to do with the plausibility of describing Beth's self-government as merely *ersatz* self-government. Mele's speculation about *ersatz* self-government is a response to the fact that the dispositions that inform Beth's post-manipulation deliberations are the result of her manipulation. Perhaps the thought is that Beth cannot undertake genuinely self-governed action after the manipulation because her *real* values are those she possessed prior to manipulation.

It seems to me, however, that this cannot be right. One of the things we are trying to do when we ask after the manipulation whether "Beth" is responsible for her actions is to decide how to respond to the person who confronts us after the manipulation, whoever that person may be. Of course, we might also wonder whether the person who now confronts us is "really Beth" – that is, we might wonder whether personal identity has been preserved across the manipulation.[30] But the issue of personal identity is separable from many of the questions we would like answered when we wonder whether it is reasonable to blame the person who now answers to the name "Beth" and whose devotion to philosophy has, let us suppose, led her to treat others with contempt or disregard. I suggest that in some cases the question of whether it is reasonable to blame Beth will be best answered by inquiring into whether she is capable of governing her behavior according to

---

well as related cases, in *Free Will and Luck* (New York: Oxford University Press, 2006). Mele's use, in the passage just quoted, of the phrase "practically unsheddable values" should also be noted. For Mele, a value is practically unsheddable if, given an agent's actual psychological constitution, and the way the world actually is, he cannot change the fact that he has that value – regardless of whether or not he might be able to change that value in some counterfactual scenario. According to Mele, autonomy does not require that an agent have played a role in the acquisition of her values, as long as her values are sheddable. Thus, Beth's lack of autonomy depends on the unsheddable nature of her implanted values.

[30] Addressing the question of personal identity will be important for resolving some questions about moral accountability. I might need to ask whether pre-manipulation Beth survived the manipulation if I am, for instance, deciding whether to press post-manipulation Beth to repay a debt or to keep a promise made prior to the manipulation. But the question of cross-manipulation identity would be less relevant if I were deciding how to respond to "Beth" after she has stolen something from me; the moral significance of this act may be entirely unrelated to the relation between pre-manipulation Beth and post-manipulation Beth.

internal values and judgments so that her behavior expresses interpersonally significant attitudes. We might ask ourselves, for instance, whether Beth's actions express judgments about the reason-giving status of other people's needs and interests and whether her actions indicate how she resolves conflicts between her own interests and those of others. And the facts about Beth's personal history give us no reason to deny that her actions convey this sort of interpersonally relevant information.

Now we might worry that given how Beth's values came about, they are not really "values" at all. In the same way that we can be externalists about intentional states like belief, we might be externalists about values and we might conclude that for a psychological state to count as a value, it cannot have resulted from manipulation. This is a worry we should be willing to entertain, but the historicist's claim is *not* that post-manipulation Beth does not have genuine values. The claim is, rather, that because Beth did not play the right role in acquiring the values of which she cannot now rid herself, these values are not really *her values* in the way required for moral blameworthiness. I contend, however, that regardless of how the issue of cross manipulation identity is resolved, as long as a set of values and preferences play for Beth the action-guiding and explanatory roles that values and preferences normally play, then these values cannot fail to belong to Beth, and to be expressed in her actions, in the way relevant to moral responsibility and blameworthiness.

While Beth did not play in her own case the (limited) role that most of us play in coming to possess our values, she apparently still governs herself according to a set of values, and the presence of these values explains her actions in the normal way. The playing of this explanatory role is what makes the values in question Beth's own in the sense that is relevant to assessing responsibility. It is important to note, for instance, that post-manipulation Beth's values have their explanatory power in virtue of informing her judgments about how to behave. These judgments are internal to Beth's psychology, so the values in question explain Beth's actions *from the inside*. This is very different from a case in which certain values explain Beth's actions only because they are the values of external manipulators who are directly causing Beth to act in certain ways. Beth's case is also very different from one in which a motivation explains an agent's action only because it is an overwhelmingly strong, non-rationalizing urge. Despite the way Beth's values and dispositions were imposed on her, she still looks out on the world from a particular perspective, governing her behavior accordingly, and it is reasonable for those affected by Beth's actions to take a moral interest in this perspective and in how her actions express it.[31]

[31] Scanlon takes a similar approach to the issue of self-formation: when we judge a wrongdoer, "we are asking whether it is appropriate to take his actions as indicating faulty self-governance. In order to claim that this is appropriate we need not also conclude that he is responsible for becoming the kind of person he now is. Whether this is so . . . is a separate consideration." Scanlon, *What We Owe to Each Other*, pp. 284-85.

After the manipulation, Beth's valuational system is the product of manipulation and it is the only such system to which she has access. Beth (or, if the reader likes, "Beth") is now an agent with a manipulated valuational system, but this need not make her current perspective any less the perspective of an integrated center of agency, and it does not make it any less *her* perspective. One way to put the general point here is to say that the question we should ask when confronted with a manipulation scenario is not really whether the values that an agent now has are *her* values – in a sense, values cannot fail to be those of the agent who acts on them. Rather, a more pertinent question is whether the entity who results from the manipulation is a moral agent. That is, we should ask whether her actions issue from the right sort of internal states such that they are capable of expressing interpersonally significant values, attitudes and judgments about reasons.[32] Insofar as moral blame responds to such attitudes and judgments, answers to these questions will tell us if it is reasonable to hold a manipulated agent responsible for her actions.

Taking up the perspective of those Beth might wrong also supports the view that she is potentially blameworthy for some of her actions. Suppose that because of Beth's implanted zeal for professional success she deliberately injures Ann in an effort to undermine the progress of the latter's research. How should Ann respond? The historicist might say that while it would be *understandable* if Ann were to blame Beth, it would not strictly be reasonable to do so because Beth's manipulation puts her beyond the proper reach of blame. The historicist might add that what Beth did was wrong, and that we should feel sympathy for Ann on this account. In other words, the historicist might distinguish between moral assessments of Beth's actions and Beth's accountability for these actions. But if we stop short of authorizing Ann to hold Beth accountable for her actions, then it is not clear that we view Beth as having done a wrong *to* Ann. And if we do think that Beth has wronged Ann by willfully treating her with undeserved contempt, then why can Ann not register this fact in the normal ways?

Of course, some agents act under conditions, or for reasons, that ought to undermine the impulse to blame them. For example, one agent might injure another by accident or because she was acting in self-defense. Now Beth's actions do not fit neatly into these or similar excusing categories, yet to say that Ann should forego blaming Beth seems to ask Ann to deny that she suffered a deliberate, unjustified harm motivated by another's contempt. To say that Beth is beyond the reach of moral blame asks Ann to view her own injury as if it were the result of an accident or as if she had been mauled by a wild animal rather than intentionally injured by a cool and reflective

[32] Manuel Vargas makes a similar point when he notes that instances of global manipulation need not be inconsistent with the obtaining of the "*basic agential structure of responsibility*." Vargas, "On the Importance of History for Responsible Agency," *Philosophical Studies* 127 (2006), p. 363 ff.

agent. To deny Ann access to the negative reactive emotions – or to count these responses as merely understandable on Ann's part – fails to take proper account of the moral wrong done to her.

And surely even one sympathetic to historicism must be struck by the implausibility of supposing that ten years might pass after Beth's manipulation and yet she never commits acts for which she is praiseworthy or blameworthy. If this seems implausible, perhaps we should say that Beth remains an inappropriate target for blame only *"for a time."* Mele himself employs this formulation several times in his recent treatment of manipulation cases in *Free Will and Luck*.[33] But what could change over time so that a manipulated agent who is not initially responsible should become so? It is a familiar fact that children pass somehow from being non-responsible to being responsible agents, but post-manipulation Beth already has the psychological and emotional sophistication that helps to distinguish adults from children.

Given Beth's immediate contentment with how she is after the manipulation, it is hard to see how anything necessary for praise and blame could occur so that Beth *becomes* responsible. Perhaps we are inclined to suppose that, if we give Beth time, then she can step back from her new values to see if she *really* accepts them. But, of course, any later evaluation of her values is likely to be tainted by the earlier manipulation, just as was Beth's immediate acceptance of those values.[34] I contend that, if certain values play the role for Beth that values normally play in bringing about people's actions, then there is nothing Beth needs to do – no stance she needs to take, no decision she needs to make – to make those values, and her subsequent actions, more fully her own.

Now one can say all I have about Beth and still hold that the treatment she received embodies a flagrant disregard for the rights we suppose a person to have over her own physical and psychological person. David Zimmerman, for instance, says that someone like Beth "may feel cheated of something valuable, namely the opportunity for naturalistically realized self-creation."[35] I admit that Beth has lost a valuable opportunity, but this does not mean that access to this opportunity is essential for moral responsibility. Indeed, it is

---

[33] Mele, *Free Will and Luck*, pp. 179, 180 and 183-84.

[34] An anonymous referee for this journal has suggested that while features of Beth's implanted psychological framework are initially unsheddable (see note 29 for an account of "sheddability"), "it might be the case that such structures ordinarily become sheddable [over time], and thus, on Mele's view, the kind of thing for which one can become responsible." It seems to me, however, that even if aspects of Beth's implanted psychological framework become sheddable over time, her decisions about whether to actually shed these values may still be affected by the brainwashing to which she was subjected, in which case her responsibility for failing to shed these values should still be called into question on historicist grounds.

[35] Zimmerman, "That Was Then, This is Now," p. 649. Michael McKenna also discusses the sense in which Beth was wronged and how this realization might affect our responses to her. McKenna, "Responsibility and Globally Manipulated Agents," *Philosophical Topics* 32 (2004), pp. 183-84.

curiously difficult to precisely characterize the opportunity that Beth lost, or to explain why it is valuable. For one thing, Beth lost the opportunity to make choices that would bring about her self as it actually is after the manipulation. But perhaps this is not a great loss since she came to be that way anyway. Beth also lost the opportunity to bring herself about as a person different from the person she actually became. However, from Beth's current perspective, this also does not seem a very great loss since Beth is pleased to be the sort of person she is and does not want to be different.

What makes Beth's manipulation so offensive is, of course, that her dean's actions followed from his own responses to reasons. The dean apparently judged that the objections Beth would have raised to his plan were insufficient to override the gains he hoped to achieve by undertaking the manipulation. Had Beth's post-manipulation state been brought about in another way (by a stroke, for example), then we would not view that state as the result of a moral infringement.[36] But while the moral character of Beth's present condition depends on how it was brought about, our assessment of Beth's blameworthiness need not.

V. Compatibilism and Ultimate Responsibility

I have tried to indicate why, on one plausible compatibilist picture of what moral blame is about, it would be reasonable to target someone like Beth with the negative reactive attitudes that characterize blame. If a reader is uneasy with this conclusion, this may be because, even though post-manipulation Beth acts purposively and reflectively, she is also an obviously passive recipient of important features of her psychology. When we focus on this passivity, it can seem inconsistent with deep moral assessment and accountability. But it is important to note that science-fiction manipulation examples, like the one about Beth, emphasize a passivity that may ultimately characterize everyone's acquisition of values. The difference is that whereas Beth's passivity is obvious in Mele's example, it is very easy for people to ignore how their own values result from factors beyond their control and how their own exercises of agency may be ultimately accounted for by reference to factors with respect to which they are passive.

Compatibilists, at least, are willing to allow that all instances of human agency might issue from a larger network of impersonal causes. But this possibility raises the worry that even when we take ourselves to be most active and to be most fully the originators of effects in the world, we are still, from another perspective, passive conduits of external causal influences. Under the

[36] Making a related point, Nomy Arpaly notes that, "Anyone who wishes to argue that Beth is not morally responsible for her actions would need to explain why having been influenced by an evil human being exempts from responsibility in a way that having been influenced in a similar way by some unlucky chance of a force of nature does not." Arpaly, *Unprincipled Virtue: An Inquiry Into Moral Agency* (Oxford: Oxford University Press, 2003), p. 129.

pressure of this perspective, our agency may seem to shrink to an "extensionless point," as Nagel puts it in "Moral Luck."[37]

Libertarian responses to the sort of worry Nagel evokes often involve attempts to distinguish at least some instances of human agency from an otherwise encompassing causal network so that at least *that* activity cannot be accounted for by factors over which the agent has no control. This is what it takes, says the libertarian, to secure *genuine* moral responsibility, and the libertarian may claim that the compatibilist's failure to make a similar move leaves her with an impoverished account of responsibility: an account on which even the manipulated characters in B. F. Skinner's *Walden Two* would count as morally responsible.[38]

David Zimmerman says that the debate about historicism is "a struggle for the soul of compatibilism."[39] One way of drawing the battle lines is in terms of how compatibilists respond to the libertarian criticism just mentioned. One response is to add historical conditions to a compatibilist account of responsibility.[40] Instead of taking this conciliatory approach, I believe that compatibilists should work harder to make it clear how an account of responsibility is not substantially weaker for being ahistorical.

One reason for compatibilists to take a hard line here is that even a historicized compatibilism seems to have difficulty describing an agent whose values are not, ultimately, a product of factors beyond his control. If this is so, then it is not clear that the addition of historical conditions to compatibilism really achieves anything. To see this point, consider the very first manipulation example I mentioned above – Mele's Beth/Manson case – or, rather, consider a recent presentation of this case by Mele, in which Beth becomes a psychological twin, not of Charles Manson, but of a similar figure named Chuck.

Chuck is unrepentantly evil. He possesses, and willingly acts on, what we might call Mansonian dispositions, deliberately injuring others for his own pleasure and selfish purposes. Beth is a very different person:

> When she crawled into bed last night, she was a sweet person, as she always had been. But she awoke with a desire to stalk and kill a neighbor, George. . . . What happened is that, while she slept, a team of psycholo-

[37] Thomas Nagel, "Moral Luck," in T. Nagel, *Mortal Questions* (Cambridge: Cambridge University Press, 1979), p. 35.

[38] To see how a criticism of compatibilism along these lines is motivated, see Robert Kane's discussion of "covert nonconstraining control," which Kane says is the kind of control that the planners of Walden Two exert over its residents. Kane, *The Significance of Free Will* (New York: Oxford University Press, 1996), pp. 64-9.

[39] Zimmerman, "That Was Then, This is Now," p. 648.

[40] I suspect that compatibilists – like Fischer and Ravizza – who take this route are led by intuitions that are better suited to an incompatibilist framework. However, Manuel Vargas counsels caution in ascribing this motivation to historicizing compatibilists. Vargas, "On the Importance of History for Responsible Agency," pp. 360-61.

> gists that had discovered the system of values that make Chuck tick implanted those values in Beth after erasing hers.[41]

According to Mele, after the manipulation, both "Chuck and Beth satisfy Frankfurt's conditions for being morally responsible"; they both act based on an integrated set of values that they can examine from a critical perspective.[42] And when Chuck and Beth take up this critical perspective, they both wholeheartedly endorse the values and desires that motivate their actions.

Chuck has never been subjected to the sort of manipulation to which Beth was subject, so, while Chuck is (*prima facie*) responsible for his behavior, Mele suggests that Beth is "too much a victim of her manipulators to be morally responsible . . . . whereas Beth exercised no control in the process that gave rise to her Chuckian system of values and identifications, Chuck apparently exercised significant control in fashioning his system of values and identifications."[43]

Here are the relevant details of Chuck's history:

> "[Chuck] enjoys killing people, and he is wholeheartedly behind his murderous desires . . . . When he was much younger, Chuck enjoyed torturing animals, but he was not wholeheartedly behind this. These activities sometimes caused him to feel guilty . . . . However, Chuck valued being the sort of person who does as he pleases and who unambivalently rejects conventional morality . . . . He intentionally set out to ensure that he would be wholeheartedly behind his torturing of animals and related activities . . . . His strategy worked."[44]

Presumably, some cold-blooded killers are responsible for their actions even if they did not take conscious steps to ensure that they would become cold-blooded killers. So the steps Chuck took are not a necessary condition on being a morally blameworthy killer, but if anyone satisfies a sufficient condition on being a blameworthy killer, then Chuck does.

However, an important question remains. Chuck took steps to make himself (more) evil. But what kind of person would choose to make a cold-blooded killer out of himself? Other things being equal, the answer is that only someone who possesses Chuckian/Mansonian values in an incipient form – someone to whom life as a cold-blooded killer seemed choice-worthy – would make the deliberate choices that Chuck made.[45] But if we are to hold

---

[41] Mele, *Free Will and Luck*, p. 171.
[42] Mele, *Free Will and Luck*, p. 172.
[43] Mele, *Free Will and Luck*, p. 172.
[44] Mele, *Free Will and Luck*, p. 171.
[45] Gary Watson expresses a similar thought about Robert Alton Harris, who I mentioned briefly in this paper's introduction. Watson notes that even if we thought that Harris had intentionally "launched himself on his iniquitous career, we would be merely postponing the inquiry, for the will which could fully and deliberately consent to such a career would have to have its roots in a self which is already morally marred." Watson, "Responsibility and the Limits of Evil," p. 250.

Chuck responsible for his adult behavior, should we not now ask if Chuck's early possession of proto-Mansonian values also resulted from an exercise of Chuck's agency? The historicist intuition is that to be responsible for acting on his bad values, Chuck must be responsible for becoming Chuck. But even if a younger Chuck did make his choices with the explicit aim of becoming just the way Chuck actually turned out, why should our historical enquiry end there? Why not also require that, for Chuck to be responsible for becoming Chuck, he must be responsible for the fact that it seemed choice-worthy to him to become Chuck? This enquiry cannot go on forever, yet the historicist can give no reason for calling it off at any particular point.

Ultimately, Chuck simply found himself to be a certain way and found himself to be satisfied with being that way. It is unfortunate for him – and for us – that he found himself to be as he is, but he is just as powerless over his initial constitution, and over the fact that he desires to change himself for the worse, as Beth is powerless over how she was manipulated. This is a function of the fact – if it is a fact – that all our exercises of agency are potentially accountable for in terms of factors over which we have no control. And this is a situation with which the compatibilist must ultimately reconcile herself. To insist that compatibilism should include historical requirements indicates, I think, a failure to fully accommodate the demands of compatibilism.

Conclusion

I have treated moral blame as justified when one person correctly judges that another has guided his actions in a way that expresses contempt or disregard for the first person's moral standing. The justification of blame, then, has mainly to do with our capacity to guide our actions in a way that reveals our attitudes toward others, and this requires no investigation into how a wrongdoer came to possess the dispositions that incline him to exercise his power of self-governance as he does. Thus, if Beth/Manson were to injure someone maliciously (yet calmly and reflectively), it would not be inappropriate for the injured person to claim that the action constitutes an unjustifiable and blame-grounding rejection of his moral standing, even though Beth/Manson played no role in becoming the way she is.[46]

Matthew Talbert
West Virginia University
Department of Philosophy
Matthew.Talbert@mail.wvu.edu

[46] I would like to thank Gary Watson, John M. Fischer, Pamela Hieronymi, Andy Reath and Dana Nelkin for their helpful comments on drafts of this paper and for their discussions of its themes with me.