

Compatibilism: The State of the Art

Michael McKenna

Those interested in the state of the art are invited to read this supplementary section. It surveys six of the positions currently on stage. Each view will receive only a brief discussion. [For other survey-oriented discussions of the state of the art for compatibilism, see the fine essays by Berofsky (2002), Haji (2002), and Russell (2002b) in Kane, ed. (2002).] More advanced treatment is best found in the current journals or directly from the relevant author's book-length treatments.

Before discussing any particular view, it is worth focusing on a troubling question over which contemporary compatibilists are in disagreement. Frankfurt's (Section 5.3), Wolf's (Section 5.4), and Fischer's (Section 5.5) views each face problems with manipulation cases. Recall, in a manipulation case, an incompatibilist opponent puts before the compatibilist a case involving an agent who satisfies all of the compatibilist-friendly requirements—hierarchies, meshes, or mechanisms—but who comes to satisfy those requirements through a process of manipulation that intuitively suggests that the agent is not free and morally responsible. The incompatibilist uses the examples 1) to force the compatibilist to acknowledge that it matters how it is that an agent comes to have the relevant compatibilist-friendly characteristics, and 2) to challenge the compatibilist to show what relevant difference there is between an agent who is manipulated into the relevant state, as opposed to an agent who has come to be in that state by a typical deterministic history.

It seems that the compatibilist has one of two means of responding to manipulation cases. One approach is to resist intuition and maintain that agents so manipulated can act of their own free will and be morally responsible for their conduct just as determined agents can. This is to bite a fairly large bullet. But some contemporary compatibilists have taken this route. Another approach is to show that the manner of acquiring the relevant compatibilist-friendly structures does matter, and that some causal sources or histories giving rise to agents are freedom and responsibility undermining while others are not. Of course, the compatibilist will try to show that determinism is consistent with a history that is not freedom or responsibility-undermining.

A very plausible way to make the divide between the two compatibilist strategies above is to distinguish between *internalist* theories of free will and moral responsibility on the one hand, and *externalist* theories of free will and moral responsibility on the other (For example, see Mele 1995; and Fischer and Ravizza, 1998). Internalist theories look only to internal features of an agent's psychological resources as they pertain to the production of action. They acknowledge that any manner of acquisition of the relevant psychological structure or mesh is irrelevant to whether at a moment in time it provides an agent with a freedom and responsibility-conferring structure. Internalist theories are thus *current time-slice theories*. They require only the satisfaction of a certain psychological structure at a moment in time. In this way, internalist compatibilist conditions for free will and moral responsibility are comparable to other *current time-slice properties* such as weight or shape, neither of which are impugned by the manner in which an object comes to have the weight or shape it has. These sorts of properties are often called *snapshot properties*. Externalist theories consider external features of an agent's psychological resources as they pertain to the production of action. They acknowledge that the freedom and responsibility-conferring psychological structure sought at a moment in time is partially determined by the causal history giving rise to that structure at that moment in time. That history will include casual features stretching outside the agent herself, such as the manner of education and socialization that shapes a person's values and beliefs. Externalist theories are thus *historical theories*. They require a sort of history that is freedom and

responsibility-conferring as opposed to freedom and responsibility-threatening. In this way, externalist compatibilist conditions for free will and moral responsibility are comparable to other *historical properties*, such as a sunburn or being a genuine dollar bill, each of which requires a certain causal history: If a burn is not caused by the sun, then it cannot be a sunburn; if a dollar is not produced by the relevant government sanctioned means, then it is not genuine but counterfeit.

In a highly influential paper, “Responsibility and the Limits of Evil: Variations on a Strawsonian Theme,” the compatibilist Gary Watson took a very honest look at how historical considerations influence intuitive judgments of moral responsibility (1987). While entertaining the internalist position, Watson proceeded to call attention, in forceful detail, to just how much historical considerations *do* matter. Watson focused upon a case of extreme evil, a horrible crime by Robert Alton Harris for the murder of two young boys. Drawing upon a couple articles from the *Los Angeles Times*, Watson first quoted a description of Harris's crime and his attitude afterwards. The story is sickening. Harris's was an act of astounding callousness. But Watson then quotes from a follow up article in which Harris's life was detailed, a life punctuated almost exclusively by abuse and cruelty, youth detention centers and then adult prisons. Watson is careful to avoid the suggestion that Harris' history caused him to be in some way incapacitated for morally responsible agency. Indeed, on an internalist, compatibilist-friendly account, it looked as if Harris had a terrible history that caused his being a free and morally responsible agent—an especially morally contemptuous one. But Watson observes that a recognition of Harris' past gives one reason to regard with ambivalence judgments of blame. Somehow recognizing a history that shapes a person to be terribly evil can influence our judgments of a person's moral responsibility for even the most evil of conduct.

Watson's essay, and the case he featured, the case of Robert Alton Harris, have had a powerful influence on recent compatibilist work. The compatibilist does have a burden here, and Watson makes this clear. How is it that histories matter? Even the internalist should have something to say in response to the appearance that histories do matter. Of course, the internalist's response will have to show that when histories do matter, they aid in understanding a freedom and responsibility-undermining impairment that itself involves only snapshot properties.

A. Identification within a Hierarchical Theory

As discussed above (in Section 5.3), Frankfurt's hierarchical theory needed supplementing so as to avoid the problem of an ever ascending conflict of higher-order desires. Recall that, one way an agent's freedom might be jeopardized is when she faces conflicting first-order desires. Frankfurt theorized that an agent's freedom (of the sort required for moral responsibility) consisted in her being able to form second-order volitional desires through which she would *identify* herself with one desire and distance herself from another. (This formed the basis for categorizing Frankfurt's as a real self theory.) But this hierarchical step is not alone sufficient to capture the relevant freedom since an agent could also face conflicts at the level of second-order, or even higher-orders of desire.

The viability of Frankfurt's hierarchical model is in jeopardy. His account of identification needs buttressing. Frankfurt's recent work on the topics of freedom and the will, as well as different features of personhood (such as the emotion of love), bolster his original appeal to identification, an appeal which previously rested solely on an agent's *having* a higher-order volitional desire that some first-order desires be effective in leading her to act. One effort Frankfurt made to speak to the problem of conflicting higher-order desires is to hold that a person properly identifies with a lower-order desire (normally a first-order desire) when she has an *unopposed* higher-order volition to act in accord with the lower-order desire, *and* she judges that further deliberation would not influence her resolve (1987). But suppose that a person's unopposed higher-order desire, the one on the basis of which Frankfurt's theory grounds the person's identification, is one towards which the agent is *passive*. The person's will is effected by this desire as an external one that has, so to speak, merely “happened” to the agent. (Consider the case of a willing addict who has simply given up the struggle and drearily resigned herself to her addiction. She'll be willing; why resist? There is nothing she can do to stop the train she is on!) Such a person would satisfy Frankfurt's amended theory, but still seem not to identify in a deep sense with her will, not in a way that would reveal her real self. So Frankfurt further amends his view with the requirement that an agent's identification must

be *active*. Her stance towards her own will is that it *determines itself* (1994). She and it are fully integrated. When an agent's will is so fashioned, then she reveals her real self in it; she regards it *wholeheartedly*.

Perhaps Frankfurt's treatment of wholeheartedness has shielded his hierarchical view from the problem of ever ascending conflicts amongst higher-order desires. But what about Frankfurt's response to manipulation cases? Could Frankfurt's compatibilist-friendly conditions be artificially induced? In considering the prospect of adding to his account an historical condition so as to rule out manipulation cases, Frankfurt writes:

What we need most essentially to look is, rather, certain aspects of the psychic structure that is coincident with the person's behavior....

A manipulator may succeed, through his interventions, in providing a person not merely with particular feelings and thoughts but with a new character. That person is then morally responsible for the choices and the conduct to which having this character leads. We are inevitably fashioned and sustained, after all, by circumstances over which we have no control. The causes to which we are subject may also change us radically, without thereby bringing it about that we are not morally responsible agents. It is irrelevant whether those causes are operating by virtue of the natural forces that shape our environment or whether they operate through the deliberate manipulative designs of other human agents (2002).

For Frankfurt, free will and moral responsibility are snapshot properties. Frankfurt is a pure internalist. Tough-minded compatibilist that he is, he simply bites the bullet. Determined agents are (in some cases) no different than agents manipulated in other ways; still, some of them act of their own free will.

B. Moderate Reasons-Responsiveness

In recent work, John Martin Fischer joined company with Mark Ravizza. Together they addressed unsettled business left in Fischer's earlier reasons-responsive account. As set out above (see section 5.5), Fischer defends a mechanism-based, actual sequence form of reasons-responsiveness. This view allows Fischer to couple reasons-responsiveness with guidance control as the freedom relevant condition for moral responsibility. His earlier account faced two serious problems, of which he was fully aware. One had to do with a proper specification of the scope of the reasons to which a morally responsible agent's mechanism of action is responsive. Making the mechanism too responsive to reasons (via strong reasons-responsiveness) sets the bar too high. Those doing moral wrong knowingly would fall short, and hence count as not acting freely *merely by virtue of their wrongful conduct*. But making the mechanism too unresponsive (via weak reasons-responsiveness) allowed a person with only a very limited or insane pattern of sensitivity to count as satisfying the freedom condition. This set the bar too low. Another problem that Fischer's earlier account faced had to do with how an agent might come to own the reasons-responsive mechanism that does issue in her freely willed actions. This problem was connected with manipulation worries since, without some ownership condition on a reasons-responsive mechanism, Fischer's view was open to criticism by appeal to cases in which an agent acts from an artificially installed reasons-responsive mechanism. In *Responsibility and Control: A Theory of Moral Responsibility* (1998), Fischer and Ravizza advance a rich account of guidance control designed to address these two problems.

B.1 Moderate Reasons-Responsive Mechanisms

Fischer and Ravizza seek to slip between the two extremes of weak and strong reasons-responsiveness by advancing an account of moderate reasons-responsiveness (1998, pp. 62-91). Essentially, their goal is to show that an appropriately sensitive reasons-responsive mechanism responds to a rich pattern of like fashioned reasons, reasons that hang together rationally as a class and fit a coherent or sane pattern. All the same, the mechanism needn't respond to all good reasons to act otherwise. As just one example of what Fischer and Ravizza have in mind, suppose Matilda, who is at a dance, is having the time of her life and would not stop Waltzing for \$100.00, or for any number of other reasons that might be put to her. But if

waltzing Matilda were offered \$1,000.00 to stop dancing, needing the money as she does, she would stop dancing. If this reason is to aid in confirming the basic rationality and sanity that might bear on Matilda's conduct, then other like weighted reasons would have to affect her mechanism of action similarly. So suppose that Matilda would not stop Waltzing for \$1,001.00, or for any other sum of money other than precisely \$1,000.00. Then Matilda, it seems, would not act from a mechanism that was responsive to fairly modest rational constraints. Matilda would *not* act from an appropriately reasons-responsive mechanism.

If instead Matilda would respond to a coherent pattern of reasons (e.g., bribes of \$1,000.00 or higher, and other sorts of incentives), then she would act from a sufficiently reasons-responsive mechanism, *even if it was not a strongly responsive one*. That is, it is acceptable if the mechanism from whence Matilda acts is not responsive to all good reasons to so act. For instance, if Matilda had been told that her mother had been seriously injured in an accident and needed her help, perhaps Matilda would not have stopped waltzing even then. She might even have continued waltzing while recognizing herself, by her own standards of decent conduct, that she should stop. Imagine her proclaiming as she continued to waltz, "I really ought to stop and help dear old mum out, but I just am having too much fun!" Matilda's insensitivity to this sufficient reason to act otherwise would not impugn the requisite responsiveness of Matilda's mechanism of action just so long as her insensitivity to this reason is situated within a set of cases that demonstrate a rich sensitivity to some rational and stable range of reasons.

This brief description of Fischer and Ravizza's reasons-responsiveness barely scratches the surface of their account. For instance, Fischer and Ravizza distinguish between different features of reasons-responsiveness. A typical deliberative mechanism in the actional repertoire of a normally functioning agent involves elements that are *receptive* to reasons, and other elements that are *reactive* to reasons. One feature of responsiveness, the receptivity feature, allows an agent to come up with and process good reasons. Another feature, the reactivity feature, allows an agent to act upon the good reasons the agent recognizes at the receptivity stage. (In the example of Matilda and her injured mother, Matilda, by way of her mechanism of action, was receptive to the reason to stop waltzing and help her mother. But she was not reactive to it.) Fischer and Ravizza also require that the mechanism be responsive to some moral reasons.

Several intricate debates have already emerged in discussion of Fischer and Ravizza's account of a moderately reasons-responsive mechanism (McKenna, 2000, 2001b; Mele, 2000; Russell, 2002a, 2002b; Stump, 1996b; and Watson, 2001). For instance, on Fischer and Ravizza's view, a determined agent can be blameworthy for knowingly doing moral wrong from a reasons-responsive mechanism because, in other worlds, were different reasons brought to bear upon the person's mechanism of action, the agent, by way of the mechanism, would act otherwise. But consider the reason to act otherwise that was present to the agent in the actual world (not doing moral wrong), the one the agent actually ignored. Given that determinism is true in the actual world, in the nearest possible world in which that very same reason is present, the agent does not respond differently to it. Hence, given that the agent is determined, it *seems* that she is not reactive to the very reason that serves as the basis for moral blame in the actual world. It is because she did not respond appropriately to the moral wrongness she herself recognized that she is blameworthy. The irony seems to be this: On Fischer and Ravizza's view, that an agent satisfies the freedom relevant condition for moral responsibility is confirmed by her reactivity to reasons other than the one that serves as the basis for blaming her in the actual world. That very reason, the one she failed to act on, because determinism is true, is one to which she does *not* react (McKenna, 2000; and Russell, 2002a). Although the consequence might be ironic, is it reason to reject the view? This is just one point of controversy currently taking shape over the details of Fischer and Ravizza's account of moderate reasons-responsiveness.

B.2 Taking Ownership of One's Reasons-Responsive Mechanisms

What about the other difficulty Fischer faced in his earlier defense of reasons-responsive compatibilism? How do Fischer and Ravizza stave off manipulation cases involving freedom and responsibility-undermining implantation of reasons-responsive mechanisms? They endorse the externalist position that only a certain sort of history will permit a reasons-responsive action to issue appropriately in freely willed actions (1998, pp. 194-206). This history, they theorize, is resistant to fabrication through artificial means.

Hence, Fischer and Ravizza treat free will and moral responsibility as historical and not snapshot properties.

According to Fischer and Ravizza, an agent's mechanism is appropriately reasons-responsive, and issues in conduct constitutive of guidance control, only if she has come to own that mechanism by means of a process whereby she *takes responsibility* for the mechanisms giving rise to her actions (1998, pp. 207-39). On their view, taking responsibility involves a subjective component. In order to take responsibility for her conduct, *an agent must see herself in a certain manner*. According to Fischer and Ravizza, taking responsibility requires that an agent recognize that her conduct is efficacious in altering the outcome of the world around her, and she accepts that the spectrum of morally reactive attitudes of others can be properly directed at her. That is, she acknowledges the propriety of members of the moral community placing moral expectations upon her that she guide her conduct within certain boundaries. Further, she must come to these beliefs through an appropriate means (and not, for instance, through deception or brainwashing or trickery in some manner or other). Fischer and Ravizza maintain that when an agent does come to take responsibility for her mechanisms of action through this means, she owns her mechanisms of action, and when they are moderately responsive to reasons, then she acts with guidance control and has satisfied the freedom relevant condition for morally responsible agency.

Perhaps the subjectivist feature of their view is subject to scrutiny. It entails that a person could be exempt from moral responsibility merely by failing to adopt the proper subjective perspective on her own conduct. Often times, a failure to take responsibility for one's conduct, a failure to see oneself as the source of moral harm, is precisely the basis for one's responsibility and guilt. Fischer and Ravizza defend their subjectivist component against this sort of objection, and indeed, their defense might withstand the heat put to it. The more important question, however, is whether they are able to make their ownership condition immune to manipulation cases. The crunch comes with their requirement that an agent must come to have the beliefs about herself through appropriate means. Whatever those means are, is it genuinely impossible that they could be artificially induced (McKenna, 2000; Mele, 2000; Stump, 1996b; D. Zimmerman, 2002)?

Is Fischer and Ravizza's account of owning one's reasons-responsive mechanism adequate to insulate their externalist theory from manipulation cases? If so, it appears that they have assuaged (some) source incompatibilist worries about the origins of an agent's actions. As with their treatment of a moderately reasons-responsive mechanism, controversy has begun to emerge around Fischer and Ravizza's externalist, and historical account of guidance control, an account that appeals to the subjective conditions involved in an agent's taking responsibility for the springs of her actions (See especially, Eshleman, 2001).

C. A Desire-Based Reasons-Responsiveness Theory

Ishtiyaque Haji defends a form of reasons-responsive compatibilism that differs in certain respects from Fischer and Ravizza's. In his book *Moral Appraisability*, Haji shares with Fischer and Ravizza a rejection of regulative in favor of guidance control as applied to judgments of moral responsibility (1998, pp. 16-41). Haji also shares with Fischer and Ravizza a mechanism-based reasons-responsive approach that holds fixed in analysis the springs of action actually operative in bringing about a freely willed action (1998, pp. 65-85). What, according to Haji, is held fixed in analysis is the motivational precursor of an action (a volitional desire), along with an agent's evaluative scheme. Fischer and Ravizza, unlike Haji, do not provide the sort of specificity Haji provides as to what in the actual sequence is held constant. They merely speculate that some subset of an agent's psychological characteristics plays the causal role in bringing about action. Fischer and Ravizza simply call that—whatever it is—the mechanism of action. On Fischer and Ravizza's view, it is that that gets held constant in analyses designed to demonstrate responsiveness to reason. Haji, on the other hand, gives content to what is held constant.

Haji's mechanism-based, reasons-responsive analysis of guidance control (Haji calls it volitional control) appears to have a theoretical advantage over Fischer and Ravizza's. In specifying the mechanism giving rise to freely willed action, Haji fixes upon some psychic features of agency that allow one to speculate about what flexibility one can place upon *the* mechanism while still assuming that it remains the same mechanism

in different thought experiments. But Fischer and Ravizza, relying only upon examples and intuitive treatments of them, have no principled basis for mechanism individuation.

To bring into relief how Haji's identity constraints on a mechanism of action differ from Fischer and Ravizza's, consider the following problem faced by Fischer and Ravizza. In speculating about whether a mechanism is reactive to reasons, Fischer and Ravizza imagine an example in which an agent's mechanism reacts differently only when an agent gets "considerably more energy or focus when presented with a *strong* reason to do otherwise" (1998, p. 74). Their example involves a person, Brown, addicted to Plezu. Rather than conclude that this sort of case confirms that the same sort of mechanism might possess the capacity to react to just some range of reasons, Fischer and Ravizza conclude that, as the agent responds differently only to reasons of a certain strength, then it must be that a different mechanism is at work, a mechanism different from the one that is *not* reactive to (weaker) reasons. The difficulty with this response for Fischer and Ravizza is that it appears that there is simply no basis for interpreting the imagined data as they do (McKenna, 2001b). What speaks against simply saying that the very same mechanism possesses a general capacity to react differently *only to reasons of a certain strength*? Fischer and Ravizza want to avoid this latter interpretation because it counts against their thesis that all that is required for demonstrating reasons-reactivity is that an agent's mechanism react differently to any one reason. A single case, Fischer and Ravizza hold, confirms a general capacity to react differently to *any* reason to which the agent's mechanism of action is receptive. But if it turned out that the mechanism of action seemed only to react to reasons of a certain strength, then it would undermine the thesis that the mechanism possesses the general capacity to react to *all* reasons put to it. Instead, in light of its merely reacting to a few strong reasons, it would seem only to reflect a general capacity to react to strong reasons, not any reasons. As Fischer and Ravizza acknowledge (1998, p.73), this would in turn create problems fitting their reasons responsive theory for compatibilism.

An advantage of Haji's mechanism-based reasons-responsive account is that his offers some theoretical basis with which to handle the sort of thought experiment that poses problems for Fischer and Ravizza's view (McKenna, 2001a). The proximal desire (the mechanism) that Haji seeks to hold fixed in assessing reasons-responsiveness is fixed on Haji's view by a motivational base and an evaluative scheme. Drawing upon Alfred Mele's work, Haji explains that a motivational base of any desire contains positive and negative elements that figure in the motivational strength of that desire. These factors comprise the total motivational base of a desire. On Haji's view, a desire is the same desire operative in different scenarios if one holds fixed the motivational base of the desire. Applied to cases of addiction, Haji's identification of a mechanism of action should get Fischer and Ravizza the result that they want. A severe addict *will* act from a different sort of mechanism than the sort a normally functioning agent will act upon. The general capacities to react will be impaired. An addict's volitional desire, given such a strong motivational base for the drug, will react only to certain sorts of reasons (such as fear of immediate death). This sort of mechanism of action will not be sufficiently reactive to the domain of good reasons to conclude that the agent acts freely when she acts from such a mechanism. Of course, *there is absolutely no reason that Fischer and Ravizza cannot see Haji's as a friendly amendment to their approach.*

Also, like Fischer's and Ravizza's, Haji's compatibilism is externalist. But Haji places the historical requirement in a different place from where other theorists such as Fischer and Ravizza do. For Fischer and Ravizza, the historical constraint on moral responsibility comes as a component of the freedom or control condition for morally responsible agency. So, for Fischer and Ravizza, guidance control requires the satisfaction of two conditions, that the mechanism of action is moderately reasons-responsive, and that an agent went through a process whereby she took responsibility for her mechanism of action (the latter is where the historical properties come in). But Haji's treats a reasons-responsive mechanism of action as sufficient to satisfy the freedom or the control condition necessary for moral responsibility. This seems only to require internalist, snapshot properties. Instead, Haji maintains that a condition for morally responsible agency distinct from the freedom or control condition is where an historical element emerges. According to Haji, a morally responsible agent must act freely from an authentic evaluative scheme, a scheme that is genuinely hers, and not one forced upon her by indoctrination, or electronic implantation, or by other means of manipulation (1998, pp. 124-39).

Haji's authenticity condition, or some might say autonomy condition, points to a different place than freedom or control where determinism could pose a threat to morally responsible agency. On Haji's view, manipulation cases do not threaten moral responsibility by threatening free will. As regards free will or control, Haji's position is internalist. Rather, manipulation cases threaten moral responsibility by threatening an agent's authenticity as an evaluator of what courses of action and way of life are valuable or worth pursuing. Hence, Haji's historicism is designed to fight manipulators that artificially craft an agent's framework of values.

According to Haji, authentic as opposed to inauthentic evaluative schemes arise through means that facilitate an agent's ability to evaluate or assess them, and especially, allow her freedom not to act in accord with them. For example, the sweet child whose sweetness is beat into her does not act from an authentic evaluative scheme, but one forged with no regard to foster understanding or sensitivities that allow maturation. Indeed, such forms of indoctrination retard the capacity for self-evolution. Hence, Haji's historicism, as applied to authenticity, is meant to rule out manipulation cases that implant in control subverting ways, an agent's framework of evaluations. But one might wonder, couldn't a manipulator manipulate in a manner that facilitated the relevant authenticity-friendly agential abilities. If so, would this sort of case undermine an agent's status as a morally responsible agent?

D. Normative Standpoint Compatibilism

One powerful element in Strawsonian compatibilism concerns Strawson's emphasis on the point of view of those in the moral community holding agents morally responsible (see Section 5.6). Hence, Strawson emphasizes the morally *reactive* attitudes. On one interpretation of Strawson's position, the best case for compatibilism begins from the vantage point of the reasons for holding and not holding an agent morally responsible. This vantage point looks first to the normally functioning interpersonal relationships of the members of the moral community. These relationships give shape to the moral expectations and demands within which a morally responsible agent must operate. In its strongest form, this Strawsonian approach leads one to regard with suspicion compatibilist efforts to capture free will by starting with action theoretic features of agency, features involving the springs of action (e.g., reasons-responsive mechanisms, or hierarchical or other structural features involving elements of an agent's psychology). This tactic moves *from* action theoretic consideration *to* full accounts of morally responsible agency, *and then*, to the propriety conditions for others holding an agent morally responsible. Given a certain Strawsonian orientation, the proper strategy is to begin with the standpoint of the members of the moral community holding an agent responsible. On this approach, the debate is best cast as a normative one about the sorts of principles that bear on when we should or should not treat a person as a morally responsible. From these normative questions, we can *then* extrapolate what sort of agent theoretic characteristics are needed to live up to these normative expectations. [Of course, this normative standpoint strategy, *when developed in its strongest form*, is specious. Why should it be assumed that if we begin by looking at the standpoint of holding responsible, the philosophical controversy will amount to a merely normative one? Might not that standpoint itself make certain metaphysical presuppositions about the sorts of beings seen as the appropriate objects of our normative demands? Furthermore, if the legitimacy of the standpoint of holding responsible were itself challenged, one quite natural way to defend it would be to argue that it is designed to respond to beings of a certain sort. This would naturally realign the order of priority, giving privilege back to questions of agency.]

R. Jay Wallace is the rightful heir to this Strawsonian strategy. In his extremely influential *Responsibility and the Moral Sentiments*, Wallace made this Strawsonian strategy his own, supplementing it so as to avoid standard worries associated with Strawson's formulation of compatibilism, and working towards an account of compatibilism that requires only guidance and not regulative control. Wallace gives structure to the morally reactive attitudes by distinguishing them from other attitudes and emotions. According to Wallace, a reactive attitude of resentment or moral indignation has as its object a certain sort of belief. The belief involves the judgment that a person has violated an obligation to which she should be held (1994, pp. 33-40). [By giving an account of holding responsible in terms of obligations, Wallace puts a Kantian spin on his account. For an alternative Humean approach to Strawsonian compatibilism, see Paul Russell's excellent, *Freedom and Moral Sentiment* (1995).] To hold an agent morally responsible (and blameworthy)

for some bit of conduct is to respond to her (or to believe that it would be appropriate to respond to her) with the morally reactive emotion of resentment or moral indignation.

Wallace's approach gives content to the morally reactive attitudes, thereby showing how they can be subject to critical evaluation by objective standards. If, for instance, it turns out that the belief serving as a basis for a morally reactive attitude is false, that is, if the person in question did not in point of fact violate the obligation in question, then the rational basis for the attitude is shown to be undercut. Hence, the attitude should be forsworn. Given this explication of the morally reactive attitudes, Wallace next turns to the Strawsonian question of when excuses or exemptions are appropriate. [Excuses involve specific pleas that one is not responsible for some bit of conduct, such as, "I did not see you there." Exemptions involve pleas that a person is not competent to function as a morally responsible agent, such as, "She did not understand what she was doing. She is severely mentally ill."] According to Wallace, that question turns upon two others. In the case of excuses, the question turns upon whether the agent in fact *did* violate the obligation to which others hold her (1994, pp. 118-53). In the case of exemptions, the question turns upon whether the agent possesses the general capacities to understand and act upon the moral demands placed on her by such obligations (1994, pp. 154-94). In each case, Wallace maintains, it is a normative principle of fairness that informs our excusing or exempting practices. In the case of excusing practices, the principle is one of desert: One does not deserve to be blamed for violating a moral obligation that she did not violate. In the case of exempting practices, the principle is one of moral reasonableness: It is unreasonable to demand of a person that she comply with moral demands if she simply hasn't the capacities or resources to do so.

Wallace proceeds to argue that neither of these moral principles is threatened by determinism. Determinism would not show that no one ever violates moral obligations; nor would it show that everyone is incapacitated to understand or comply with the demands involved in moral obligations. Furthermore, as Wallace observes, excuses that appear to generate the demand for alternative possibilities (and hence, Garden of Forking Paths freedom and regulative control), such as, *I could not have done otherwise*, are excuses that cannot be generalized. They work, when they do, *only* by showing that an agent did not violate a moral obligation in acting as she did (1994, pp. 152-3). Factors other than her disregard for a moral obligation explain her action (e.g., she was forced to do something at gunpoint, or was physically unable to get to the emergency phone because her leg was broken, etc.). So it seems that Wallace shares with Frankfurt and Fischer the view that regulative control is not required for moral responsibility, only guidance control is.

But where does Wallace stand as regards manipulation cases and the internalist versus externalist (snapshot versus historical properties) approach to free will and moral responsibility? In *Responsibility and the Moral Sentiments* Wallace addresses manipulation cases under the rubric of systematic behavior control or conditioning (1994, p. 155). His principle of distinguishing those exempt from morally responsible agency is informed exclusively by whether the persons in question are incapacitated to understand and comply with the demands of moral obligations. Therefore, it looks as if Wallace is committed to an internalist position, treating morally responsible agency as a snapshot property. (Though Wallace himself never discusses the debate in these terms.) To the extent that any internalist position is susceptible to manipulation cases, it seems that Wallace's is too. It looks as if Wallace must adopt that same style of response as Frankfurt's.

Wallace approaches the free will issue from the normative standpoint of moral philosophy, whereas a view like Frankfurt's or Fischer and Ravizza's approaches the topic from the standpoint of metaphysics and action theory. There appears to be a serious tension between the two approaches, and Wallace openly considers this tension, arguing that his is to be preferred to the other approach (1994, pp. 85-95). [This contrast between metaphysical and normative might well be overdrawn. For instance, on Fischer and Ravizza's view, a moderately reasons-responsive mechanism must be receptive to at least some moral reasons. Hence, "normative" considerations bleed into their metaphysical account. Similarly, on Wallace's view, the normative principles of fairness that inform our practices of holding morally responsible concern persons. Hence "metaphysical" considerations bleed into his normative account.] But any tension of methodology should not be too quickly associated with a tension in outcomes. Perhaps a view like Wallace's might mesh with a view like Frankfurt's or Fischer and Ravizza's, so that the same sort of conditions of agency and responsibility apply. This would suggest that the different approaches can actually

work in tandem, maybe because our normative principles of fairness informing our morally responsibility practices track our metaphysical and action-theoretical parsings.

E. Practical Standpoint Compatibilism

Hilary Bok has recently advanced a form of compatibilism that shares certain features with Daniel Dennett's. In particular, Bok's, like Dennett's (see section 5.2) is a viewpoint, or as she puts it, a standpoint argument. Both maintain that a certain standpoint is a legitimate one, and is the one whence judgments of responsibility arise. Responsibility or freedom concepts at work within this viewpoint, each argue, are compatibilist-friendly ones, unthreatened by the possibility that determinism is true. Bok's, however, differs in various respects from Dennett's. Unlike Dennett, Bok maintains that the relevant standpoint arises by distinguishing, in Kantian fashion, between the practical and the theoretical points of view. The former standpoint has as its goal answering questions about how one ought to act. The latter concerns truthfully describing and explaining events as necessary results of antecedent conditions (1998, pp. 62-5). It is Bok's contention that, while both libertarian notions of free will and compatibilist notions of free will *are* found within the range of the "ordinary" concept of freedom, the one that matters as regards the free will debate is the one that a practical agent would have good reason to adopt (1998, p. 100). It is not that one, so to speak, captures the "real" truth about what the concept of free will is. It's that one is useful to agency in a way that the other is not. Hence, Bok seeks to settle the free will problem by looking at the role of the concept of freedom from the point of view of a practical deliberator engaged in settling questions regarding how to act, how to live her life, what kind of person she wishes to be, etc.

Given her practical standpoint approach, Bok maintains that the sort of freedom that is of use to practical deliberators concerns possibilities restricted in scope to those consistent with what an agent understands to be practically possible from her limited epistemic perspective (1998, p. 108). These possibilities are much looser than the sort required by libertarian free will, the latter requiring attention only to possibilities given a precisely specified past and holding fixed the actual laws of nature. Bok's favored possibilities allow an agent to reason about alternative courses of action conditional upon her choosing in one manner as opposed to another. Hence, Bok embraces a conditional analysis of free will. Naturally for Bok, for an alternative to be genuine, it need not be open given precisely the same past and laws of nature. It need only be genuine given the courser facts about the conditional relation between an agent's will and subsequent conduct likely to arise from it. If she chooses in one fashion, then a certain course of action will come to fruition; if she chooses in another, then another course of action will come to fruition (1998, p. 120).

So Bok embraces regulative control in order to capture the freedom relevant condition for moral responsibility. Her appeal to a conditional analysis of regulative control, however, leaves her open to the same sort of criticism leveled against the classical compatibilist's conditional analysis (see 3.3 above). On her view, an agent who does *y* and not *x* has free will with respect to the alternative course of action *x* if the following is true: If she chose to *x*, then she would *x*. But classical compatibilists faced a troubling sort of example that seemed to disprove any such equivalence. A person could lack free will with respect to a course of action since she might suffer from a psychological condition that made it impossible for her to choose that course of action *x*. All the same, it would still be true that, *if* she *did* choose the course of action *x*, then she would *x*. Hence, the analysis generates the result that an agent has free will in cases in which clearly she does not (Ginet, 2003; Haji, 2002; and McKenna, 2002).

It would, however, be unfair to dismiss Bok's practical standpoint compatibilist theory simply because the particular account of freedom she relies upon is suspect. It might be true that conditional analyses are doomed to failure, but other treatments of agential possibility remain. So perhaps Bok could defend some notion of regulative control *sans* the classical compatibilists' conditional analysis. The important philosophical point, on Bok's approach, is that the demands of the practical standpoint invite a looser notion of possibility than the sort that is at work when formulating precise definitions of determinism, or when attempting to fashion libertarian notions of free will. Some looser notion of agential possibility might allow a compatibilist to say that all the possibility that is needed for regulative control is something like epistemic possibility, "for all I as a practical agent know" possibility.

So set aside the (legitimate) criticism that Bok's manner of developing regulative control falls prey to the same sorts of objections leveled against the classical compatibilist. Fix instead on the strategic effort to show that the practical standpoint encourages some looser notion of possibility than the sort relevant to determinism. How does Bok justify the import of this compatibilist notion of freedom? According to Bok, it is justified by our practical interest in improving the qualities of our wills, which are reflections of our selves, fashioners of the people we will become (1998, pp. 123-66). We care, when we deliberate, to evaluate possibilities in terms of how we acted in the past and how we might improve our conduct and ourselves in the future. A compatibilist notion of freedom will help us to do the work of improving our characters and fashioning our selves. It will allow us to conceive of what it is within our general range of capacities to do, and to evaluate our options in terms which lead to our improvement.

Bok's justification for a compatibilist-friendly notion of freedom is surprisingly forward-looking. We care about the relevant notion of freedom since we care about future improvements to our wills and characters. This is in deep conflict with the spirit of approaches (such as Strawson's) that have dismissed such consequentialist sorts of justification of free will as unable to capture the intimate connection between an agent and her responsibility for what she had done. For her regret or guilt to be a genuine expression of her attitude towards her conduct and those whom she wronged, it had better be a response to what she did in the past, and not merely a vehicle for improving herself in the future. But set this sort of worry aside. Another implication of Bok's forward looking account is that it seems that she must accept an internalist view (Haji, 2002). If the practical use of freedom is best understood as forward-looking, then however an agent is caused to come into existence and have the psychological structure she has, she can regard herself as free and responsible insofar as doing so will aid in improving her will somewhere down the pike. Bok, it seems, like Frankfurt and Wallace, is an internalist.

F. The Action Theory Theory

One classical compatibilist strategy that most contemporary theorists seem to have overlooked is an austere one that attempts to make do with as little as possible beyond more primitive features of agency itself. The classical compatibilist used only blunt instruments to fashion such a view (see section 3.1). All that she seemed to draw upon in an account of agency was the notion of a desire or want, and the negative condition that it be unimpeded. But despite her impoverished resources in capturing the springs of action, her strategy was a philosophically elegant one. Simply postulate no more than the features giving rise to agency. These features need be no fancier than the sort typically at play with normally functioning human persons. There is little reason to imagine that determinism is incompatible with such agency *simpliciter*. Next, to capture *freely willed* agency, append some negative conditions that secure that sometimes that basic sort of agency can function unhindered by coercive or compulsive forces. Viola, free will! Add no more metaphysical constraints, nothing further to show how it gels with determinism. Leave compatibilism at that; simple is better. [As one subject editor pointed out, there might be a lot of metaphysical work loaded into Mele's account of agency *simpliciter*, and it might be that the controversial compatibilist details are found there and not in any further metaphysical conditions (such as hierarchical accounts of the will).]

In *Autonomous Agency*, Alfred Mele sketches a contemporary form of compatibilism that shares with the classical compatibilist the strategy of theoretical austerity: Capture free will by adding as little as possible beyond an account of mere agency. Naturally, Mele draws upon a finely tuned set of conceptual tools to capture more clearly the contours of the springs of action. He makes use of philosophy's matured understanding of action theory, an understanding to which he himself has contributed. But he attempts to append to his account of agency as little as possible in order to generate a form of compatibilism that speaks to the contemporary dialectic. Given that the theory of compatibilism that he advises attempts to build mostly from his work on the theory of action, it might be called the *Action Theory Theory*. [Mele himself is not a compatibilist. He remains agnostic between compatibilists and incompatibilists, but offers each camp a best shot at a workable theory.] On Mele's view, free will is where the action is.

Mele does not fashion compatibilism in terms of free will and determinism, but instead in terms of *autonomy* and determinism. But as Mele openly acknowledges, he understands autonomy to be amongst the freedom concepts in discussions of moral responsibility (1995, p. 4). Mele understands autonomy,

basically, as the capacity for self-governance or self-rule, and he looks to more basic features of self-control to explain the phenomena. *A word of caution:* Mele's discussion of self-control should not be assimilated with the notions of guidance or regulative control discussed by Fischer (and used in this essay to distinguish different views of free will). Fischer and others who make use of the notion of control as synonymous with free will see the relevant notion of control as all that is needed to satisfy the freedom relevant condition for morally responsible agency. But Mele does not presume that the notion of self-control he wishes to distinguish can do that compatibilist work. Indeed, as will become clear, Mele holds that the sort of self-control he wishes to distinguish, while a genuine actional feature found in most normally functioning agents, is not adequate to explain the kind of freedom at issue in discussions of moral responsibility.

What does Mele have in mind by a self-controlled agent? He has in mind the opposite of an agent who is akratic, that is, an agent who is weak willed. Mele treats self-controlled and weakly willed conduct as two sides of the same coin (1995, p. 5). Hence, he draws upon his account of weakness of will to help shed light on weak will's opposite, self-control. According to Mele, weakness of will arises when one's motivational states become misaligned with one's judgments about the best (or better) course of action (1995, p. 7). Because, on Mele's view, evaluations of things one desires can be out of line with the strength of one's desires for those things, one's best judgments sometimes are in disharmony with one's strongest desires (desires that do appear to play a causal role on action) (1995, p. 25). It is in such cases of conflict that the self-controlled as opposed to the weak willed agent is able to resist acting upon her currently strongest desire and instead act in accord with what she judges it best to do (1995, p. 80). This is possible, on Mele's view, since agents with the sort of sophistication of normally functioning persons are able to promise themselves rewards for resisting temptation, or direct their attention to less desirable features of the path more strongly desired, and generally, exploit less episodic and more stable motivation to exercise self-control. In so doing, they can bring their motivational condition into line with their "best judgments" (1995, pp. 81-83).

According to Mele, an agent can be both determined *and* exercise the actional resources to act with self-control. Hence, determinism is compatible with self-controlled agency. But it is Mele's contention that even an optimally self-controlled agent can fall short of autonomous agency (1995, pp. 121-7). Hence, Mele concludes that more has to be added to his account of self-control to get all the way to autonomous agency. Notice that, up until this point, Mele's theory appeals only to features of agency at work in the theory of action. These include such notions as best judgments, strongest desires, intentions, the ability to promise oneself rewards, the distinction between more and less episodic motivational factors, and questions about the potential conflict between best judgment and motivational desires.

How is it that, on Mele's view, self-controlled agency is *not* itself sufficient to capture autonomous agency? Because, on Mele's view, the values and principles that inform one's deliberation and conduct might be installed in a person in some autonomy (or responsibility) undermining manner. Such cases arise via various sorts of manipulation, through brainwashing or hypnosis, or even a history of rigorous indoctrination during youth. The key element in an agent's *not* having had unsheddable principles and values installed in an autonomy-undermining manner is that the causal manner in which the agent acquired them did *not* bypass an agent's critical capacities to assess the principles and values for herself (1995, pp. 166-73). In such cases, she had the opportunity to embrace or shed them. Hence, Mele postulates a *negative historical constraint* on autonomous agency: An agent acts autonomously if she acts with self-control, and so long as she was not caused to endorse unsheddable values and principles (against which she is able to evaluate her reasons for action) by means that bypassed her capacities for critical evaluation.

Notice that Mele's view differs from Fischer and Ravizza's (see Section B above) in that Mele's historical constraint is a negative one. Hence, Mele, unlike Fischer and Ravizza, offers no specific account of what kind of causal history of owning one's springs of action *is* required. Rather, Mele only states what kind of history can not give rise to autonomous (or freely willed) agency. Some will probably find Mele's historical view more attractive than Fischer and Ravizza's since their positive effort to state what sort of history is required seems to lead them to a subjectivist view. On Fischer and Ravizza's subjectivist view, an agent is not morally responsible for her conduct if she does not believe herself to be a candidate for such

evaluations. On Mele's view, an agent might fail to have the relevant beliefs, and hence refuse to take responsibility for what others would insist she should. But so long as the values and principles that she does act upon did not arise through means that bypassed her capacities to evaluate them, then she is autonomous with respect to (and morally responsible for) the self-controlled actions arising from them—no matter what her subjective view of her self might be.

Of course, others will find Mele's view less attractive than Fischer and Ravizza's since they offer a positive account and Mele does not. Without a positive account, how can one be sure that Mele's theory is not threatened by manipulation cases? Is it not possible that a very crafty manipulator could manipulate an agent's *means* of critically evaluating the values and principles that she later endorses? Mightn't a manipulator do this *rather than* bypass altogether the relevant agential capacities? Without a positive theory, it is unclear how Mele can guarantee that this is not possible. If it is, then Mele will need either to develop a positive externalist account like Fischer and Ravizza have, or instead, settle for the sort of internalist position embraced by Frankfurt (as indicated above in section 6.2).

Another important feature of the compatibilist view Mele offers is that, like Fischer and Ravizza's, it also eschews the demand for regulative control. In fact, Mele, along with David Robb, coauthored one of the most compelling defenses of Frankfurt examples in order to demonstrate that moral responsibility, and also free will and autonomous agency, do not require the freedom to do otherwise (Mele and Robb, 1998). Hence Mele, like Frankfurt, Fischer, and Wallace, avoids difficulties arising from incompatibilist attacks upon regulative control, such as the one posed by the Consequence Argument (see section 4.1).

A good deal more could be said in an effort to explain the sort of compatibilism Mele suggests, but one of the more provocative and distinctive element in his account is his strategic return to the austerity of the classical compatibilists. The Action Theory Theory gets a lot of mileage just out of action theory. It is an elegant philosophical maneuver and merits more serious attention than it has received up to this point.

[Copyright © 2004](#) by
Michael McKenna <mmckenna@ithaca.edu>