

# Autonomy and Manipulated Freedom

Tomis Kapitan

*Philosophical Perspectives* 14 (2000), 81-104

## 1. Introduction

In recent years, compatibilism has been the target of two powerful challenges. According to the consequence argument, if everything we do and think is a consequence of factors beyond our control (past events and the laws of nature), and the consequences of what is beyond our control are themselves beyond our control, then no one has control over what they do or think and no one is responsible for anything. Hence, determinism rules out responsibility. A different challenge--here called the manipulation argument--is that by allowing agents to be fully determined compatibilist accounts of practical freedom and responsibility are unable to preclude those who are subject to global manipulation from being free and responsible.<sup>1</sup>

Both conclusions have prompted a variety of compatibilist responses. To the consequence argument, some have dropped the principle of alternate possibilities as a condition on control (Fischer 1994). Others have denied that the past or the laws of nature are beyond our control (Lewis 1979), and still others have questioned the validity of the argument, specifically, the closure principles upon which it relies (Slote 1981). I have argued that the appropriate combination of such responses yields an adequate rebuttal to the consequence argument (Kapitan 1996). The manipulation argument has been challenged by those who find that responsibility is not ruled out by external controllers (e.g., Frankfurt 1988, Blumenfeld 1988), and by those who think a historical conception of autonomy provides a way of disqualifying the manipulated agent from being either free or responsible (Mele 1995, Fischer and Ravizza 1998, Haji 1998). Not satisfied with these replies, I will show how they fall short and offer an alternative account of why the manipulation argument fails to refute compatibilism.

## 2. Conditions of Responsibility

Some philosophers argue that no one is ever truly responsible for what they have done since no one is truly self-determining (Strawson 1986, chp. 16, and see Double 1991, chps. 6-7, and Waller 1998, chp. 4). As a stipulation on a technical use of 'responsible' there is no need for dispute, but there are sound reasons for thinking that in some sense the existence of responsible agents is beyond doubt. For one thing, societies are fully justified in setting standards of behavior and in applying mechanisms for maintaining those standards, specifically, educational requirements and practices of praising/rewarding and blaming/punishing people for what they do. For another, these sanctioning practices are justifiably applied only to those for whom there is evidence of being worthy of praise or blame. When the evidence is at hand, agents can legitimately be

held responsible and, to that extent, *are* responsible, even if they are not "truly responsible" in some philosophers' technical sense.

Let us speak unapologetically of agents being responsible for performances or omissions of actions, and construe responsibility for any other type of situation in terms of bringing about, sustaining, or preventing that situation. Retrospectively, an agent S is worthy of praise or blame for having performed action A at time t only if there is reason to suppose that the following conditions are met at some suitably prior time:

Intention/Foresight: Doing A at t was an intentional action of S's or a foreseeable result of some intentional action of S's.

Control: Doing A at t was under S's control.

Obligation: S was subject to a moral demand as regards doing A at t.

Analogous requirements apply for omissions. While necessary, the joint satisfaction of these conditions does not settle whether S is deserving of praise or deserving of blame. For example, though a student is obligated not to shout in class, yet exercises control in shouting intentionally, he might be exculpated if he believed that overriding factors were present. Again, Sam and Sal might separately steal \$100, an act they are similarly obligated to avoid and over which they have the same control. But if Sam steals from a motive of helping a poor person in an emergency whereas Sal steals in order to buy himself expensive cognac for impressing his friends, then, other things being equal, Sam is less blameworthy than Sal. Accordingly, to determine whether S is responsible and to what degree, one must know what particular psychological states accompanied S's action, specifically, S's beliefs about prevailing circumstances and pertinent obligations together with pro-attitudes he possessed (desires, values, preferences, etc.) that were relevant to (about, or causally influential upon) the action or its results. Schematically:

Relevant States: S's motivational and doxastic states M1, M2, . . . , Mn were the relevant states with which S did A at t.

For a given agent, action, and time, a fourth responsibility condition will be an instance of this schema. Joint satisfaction of these four conditions is sufficient for responsibility.

Some qualifications are in order. First, it is adequate evidence that the conditions are satisfied that justifies one in holding S responsible for A-ing. A mere joint satisfaction of the conditions is not sufficient for anyone's holding S responsible. Nor is it necessary given that evidence is defeasible. Second, inasmuch as the conditions provide non-question-begging criteria for judging whether S is responsible, then they must be decidable independently of assessments of S's worthiness of being praised or blamed. Third, each condition can be refined to suit the specifics of particular theories of free action and responsibility. In particular, the obligation condition may or may not be

construed as implying that S actually is morally obliged with respect to A (see Zimmerman 1988, pp. 40-46 and Haji 1998, chps. 9-10). On the other hand, either this condition or the relevant states condition should guarantee that S was in a position to have understood he was subject to that demand.

The control condition is central to, if not definitive of, practical freedom, and at least two factors are involved. The first is that of efficacy: performance of the action would result in an anticipated manner from relevant pro-attitudes (intentions, desires, values, goals, etc.), so that the agent lacks control if compelled to do the action or prevented from doing it against these attitudes. The notion of efficacy can be extended to include omissions, and to the certain consequences of performances and omissions. Some hold that efficacy must be dual by including abilities both to do and to refrain from an action according to "the determination or thought of the mind" (Locke's Essay Book II, Chapter XX, section 8). Regardless whether a demand of duality can be sustained, a second ingredient of S's control over his A-ing at t is that S presumes that both doing A at t and omitting A at t are open alternatives for him. This embodies a pair of presumptions on S's part, namely, (i) that he would perform A (of refrain from A) at t were he to undertake (intend, choose) so doing, and (ii) his undertaking A at t is, as yet, contingent--where the presumed efficacy and contingency are indexed to what he takes himself to believe (Kapitan 1991, pp. 31-33, and 1996, pp. 435-439). If S is minimally rational, then by presuming his A-ing as open relative to what he takes himself to believe, it follows that his A-ing--represented by S to himself by the first-person "my A-ing"--actually is open relative to what he takes himself to then believe. This doxastic openness of A-ing implies that S's undertaking (choice) to do A is, at the same time, doxastically contingent and, hence, provides a sense in which S takes his undertaking to do A at t to be "up to himself" or "his own" (Kapitan 1989, pp. 31-34). Satisfaction of the control condition, then, implies the agent's presumption of duality even when not matched by an ability to do otherwise given the actual circumstances. In Kantian terms, responsible agency occurs "under the idea of freedom" even if the agent is not free in a "theoretical" respect.<sup>2</sup>

It is unreasonable to require that S's control over A-ing implies S's control over all the pro-attitudes and character traits that led to his A-ing. He may be unable to think about some of these attitudes and traits, and some of them may be deeply entrenched and unshakeable features of his personality. At the same, time, higher degrees of responsibility for A-ing might accrue if S has control over some relevant attitudes, and if S is responsible for having an attitude or a trait P during interval t then S must have control over P at t or during a time suitably prior to t.

### **3. Autonomy and the Manipulation Argument**

A man who is coerced at knifepoint to relinquish his wallet may very well be in control of his behavior, and his control may be dual if he is able to refrain from complying with the thief's demand. Yet he is not acting in a way that he prefers; he is not fully self-governing and the consequent actions are not wholly "his own." An addict who acts from irresistible desires for a drug might not want to so act or to be so motivated. His autonomy is also offset by the influences that frustrate his will or prevent him from acting

in a manner he would otherwise prefer. Seemingly, S's practical freedom requires satisfaction of yet a further responsibility condition, namely, that S was autonomous in doing A.

Let us be cautious and inquire whether autonomy is implied by the previous conditions. A coerced agent is exempted from blame because the circumstance of the threat produces an obligation that overrides the usual demand to do or omit an action, namely, to spare oneself (or someone else) the threatened harm. One who is subjected to irresistible desires lacks self-control with respect to certain actions, and the same is true of akratic agents in general. Perhaps what is lacking in both cases is the agent's *identification* with the proximal psychological causes of the action, in which case a specification of the relevant states condition is what fails to be satisfied. If so, then autonomy is captured in the responsibility conditions already identified.

But suppose S were controlled by another agent who causes S to behave, think, in accord with its own desires, not by causing S to act in a manner contrary to S's own intentions or preferences, but by manipulating S's choices, preferences, and beliefs through non-coercive means such as indoctrination, information control, hypnosis, drugs, or brain implants. Oblivious to these causes, S is subject to covert, non-constraining control (CNC control) exercised by external agents (Kane 1985, p. 35). If the control is so comprehensive that all S' desires, values, deliberations, choices, intentions, etc. within a certain interval are what they are because of the manipulation, then it is global during that interval (Schoeman, 1978, p. 295). If introduced at a point after the beginning of S's history then it is interruptive, while if S's whole mental life from birth is so determined then it is total. S's psychological life might be disconnected and episodic if the manipulators are constantly inducing the states, but it might be coherent and well-integrated if manipulation occurs through successful programming (Frankfurt 1988, p. 53). Maybe Skinnerian behavioral engineers are capable of imposing this sort of global control over an individual, or perhaps total control could be achieved through devices implanted in S's brain, beginning, say, with the insertion of microchips into S's fetus, permitting the external controllers to both monitor and direct the course of S's psychological development within given genetic constraints. The implants might even be pre-programmed, allowing the manipulators to retire from the scene, confident their aims will be carried out. The picture is extreme, but advances in micro-processing and robotics that would make it possible cannot be ruled out. What seems clear to a good many observers is this: a covertly manipulated agent is much like a robot, perhaps a living, breathing, conscious organic being capable of informed rational deliberation, but not an autonomous agent, and, by that very fact, neither free nor responsible.

Manipulation undermines responsibility if it moves an agent against his will or better judgment, or if breaks psychological continuity by introducing a hiatus between an agent's current preferences and previously entrenched dispositions (Haji 1998, pp. 116-119). We might think a manipulated agent lacks autonomy and responsibility even if psychological continuity is not broken and he enjoys a coherent, well-integrated character complete with a sufficiently well-informed sense of responsibility and a high degree of

self-control. But how does he lack autonomy if he is self-controlled? Moreover, how is he any different from a "naturally determined" self-controlled agent?

The question spawns an intriguing analogical argument against compatibilism. Robert Kane asks us to consider a person, Ishmael, in a world W0 whose his entire inner life and behavior is the consequence of CNC control by external agents. Kane assumes that by being so controlled, Ishmael lacks what is required to be a responsible agent in W0. Imagine that in a similar world W1, Ishmael, or a counterpart of Ishmael, acts and thinks exactly as Ishmael in W0 does, with exactly the same capacities, abilities, and opportunities that Ishmael in W0 has, but as a result of natural forces rather than manipulation. Suppose, moreover, that Ishmael in W1 satisfies the conditions of responsibility that compatibilists have traditionally set forward, in particular, he possesses all the control over his actions and mental states that a compatibilist could conceivably propose. Since Ishmael in W0 also satisfies those conditions then he has no less control or autonomy and, hence, is no less practically free. Consequently, he too is responsible. Since this is absurd, compatibilist theories about practical freedom and responsibility are mistaken. In sum:

1. A manipulated agent is neither practically free nor responsible with respect to any of its behavior or states resulting from the actions of external manipulators.
2. There can be a totally manipulated agent who is practically free in any sense a compatibilist can offer, in particular, who can satisfy the conditions of control and autonomy to the degree that any naturally determined agent can.

Therefore,

3. The naturally determined agent is neither practically free nor responsible with respect to any of its behavior or states.

Kane concludes that since "no existing compatibilist account of freedom" can block the CNC control of a compatibilistically construed "free" agent by another agent, then all such accounts of responsibility fail (1985, pp. 37-42).

Kane classifies a compatibilist response to this manipulation argument as hard if it denies premise 1, and soft if it draws the line at premise 2 (Kane 1985, 38). The hard compatibilist takes the totally manipulated agent to be free and responsible for whatever behavior, mental states, and character were within its control. The obvious problem with this position is that we would exculpate someone from responsibility were we to discover that he was subject to total CNC control in behaving as he did, just as we would were we to discover that he was a victim of malicious indoctrination since youth or of an injury to his prefrontal cortex that radically changed all his decision processes.<sup>3</sup>

The soft compatibilist who denies premise 2 faces the task of explaining why covert manipulation diminishes responsibility and practical freedom while the natural influences to which a person is subject, viz., genetic inheritance, parental guidance, socialization, and education, do not (Kane 1996, 68, 225, n.15). One approach insists that the totally manipulated agent is not autonomous because its inner states are produced by something alien to the "self." It repudiates the notion that autonomy can be captured by purely internalist descriptions, e.g., efficacy of pro-attitudes with respect to proximal states or conformity of these states with higher-order identifications, values, or a deeper self. Instead, autonomy is a matter of the historical genesis of an agent's psychology and can only be described through an externalist account of the acquisition and retention of pro-attitudes. Obviously this approach cannot preclude these attitudes from being the product of some external influences, nor merely preclude manipulation from being among the permissible external causes. What it must offer is a non-circular explanation of why manipulation erodes autonomy whereas causation by "natural" means does not (see Dworkin 1976, p. 24; Christman 1989, pp. 10-22; Kane 1996, p. 68-69).

In what follows, two externalist accounts of autonomy are examined and rejected. It is then argued that while manipulation does not automatically rule out responsibility, cases where it does are explainable in terms of the four conditions of responsibility without invoking an independent requirement of autonomy.

#### **4. Autonomy and Historicity: Alfred Mele**

Alfred Mele acknowledges that agents who possess ideal self-control over their own motivations and decisions fail to be autonomous if they are controlled by external manipulators (Mele 1995, p. 122). In examining cases of brainwashing, he clearly sees the threat of the manipulation ("mind-control") argument to a compatibilist theory of responsibility:

How is the control that manipulators exert over an agent in the brainwashing scenarios examined here, for example, relevantly different from the effect of the distant past on an agent at a deterministic world? (Mele 1995, p. 173)

He considers two agents, Ann and Beth, who are similarly motivated by an entrenched or "unsheddable" value of being an industrious philosopher, but whereas Ann has arrived at her motivation "on the basis of careful critical reflection over many years," Beth's motivation is due to the intervention of brainwashers who have successfully altered her values and priorities. Were Beth motivated to perform wicked Charles Manson type deeds then she "is not responsible for her Mansonian character--values principles, and the like" (p. 161), and "we would not hold her responsible for her Mansonian character" (p. 159). Ann, by contrast, is responsible for what she has become. Why the difference? Ann, unlike Beth, is an autonomous agent because she has autonomously developed her values (p. 155). Lacking autonomy, a manipulated agent like Beth fails to be practically free.<sup>4</sup>

Etiology is at the core of Mele's concept of psychological autonomy. Internalist conceptions of autonomy are vulnerable to the manipulation objection even when control is dual (pp. 147-152). What a manipulated agent lacks is authenticity, the ability to subject its own drives to rational scrutiny (see also Feinberg 1989, p. 32 and Dworkin 1989, p. 61). According to Mele, an agent S does not authentically possess a pro-attitude P during interval t if S was compelled to have P in a way not arranged by S, that is, if in acquiring P, S's own capacities for control over his mental life were bypassed by external causes (pp. 166-167)--where capacities are bypassed insofar as they play no role in the acquisition (Blumenfeld 1988, 222). More fully, if (i) S comes to have P in a way that bypassed capacities for control over his mental life in a way he did not himself arrange, (ii) P is "practically unsheddable," (p. 153), and (iii) S does not have other pro-attitudes that would support his identifying with P (excepting those unsheddable ones that have themselves been induced in a bypassing manner) then S is compelled to possess P and, accordingly, is not autonomous (p. 172).<sup>5</sup>

Unlike Ann, Beth is compelled to have her values because they were induced by external agents in a manner not endorsed by her. Even if her change of character resulted from passing through a randomly occurring electromagnetic field in the Bermuda Triangle, she would still be compelled to have a pro-attitude because her capacities for control were bypassed. Similarly, a person can have compelled pro-attitudes as a result of indoctrination, for example, a child whose religious convictions resulted from brainwashing by a religious fanatic who taught that critical reflections about these doctrines will earn eternal damnation in hell. The child's capacities are bypassed either because its modest capacity to believe and desire on the basis of an assessment of evidence was circumvented, or, if this capacity had not yet emerged, because its capacity to develop into an ideally self-controlled agent was bypassed, "and, indeed, destroyed" (p. 168). Besides psychological compulsion, autonomy can also be undermined by controlling the information to which agents have access, or by preventing agents' deliberative processes from reliably resulting in intentions and actions (pp. 179-185). In the latter event, a capacity to develop into agents who have "a significantly greater role in shaping their own deliberative habits" is bypassed by the fact that unreliable or inefficient deliberative habits are engineered.

Mele offers (p. 187) the following trio of conditions as sufficient for an agent S to be psychologically autonomous:

1. S has no motivational states that are compelled or coercively produced.
2. S's beliefs are conducive to informed deliberation about matters of concern to him.
3. S is a reliable deliberator (i.e., can effectively act upon deliberations and intentions).

While not claiming satisfaction of these conditions to be necessary for psychological autonomy, Mele argues that an agent satisfying them will not be subject to CNC control despite being determined. The same holds for one who meets the necessary and sufficient conditions that a compatibilist could offer for strong psychological autonomy, namely,

regularly exercises ideal self-control, is mentally healthy, and satisfies conditions 1-3. Since manipulation entails compulsion and compulsion blocks autonomy, then the manipulated agent cannot be autonomous in a way that the naturally determined agent can. Their difference is a matter of how their respective pro-attitudes were generated. Mele concludes that compatibilism is saved by exploiting the standard compatibilist distinction between mere causation and psychological compulsion, employed not only at the level of choice and action, but at the level of "character" (p. 173).

Despite its impressive detail, Mele's account is not an adequate response to the manipulation argument. Either it makes autonomy too rare to be of much use in a theory of responsibility or it cannot block a CNC controlled agent from being autonomous. Consider the first horn of this dilemma. If an unsheddable pro-attitude is acquired in a way that bypasses capacities for self-control, say, by being manipulated, exposed to a randomly occurring electro-magnetic disturbances, or involuntarily ingesting a mind-altering drug, then the agent is compelled to possess it and is not autonomous. Suppose Beth acquired a new value in a less spectacular fashion, for instance, as a result of moving to a new locale, taking a new job, and being subjected to a different life style and set of environmental stimuli. She developed an aversion for people with high-pitched voices, a singular value that does not cohere with the rest of her standing values, though it does not necessarily contradict them either. Moreover, she did not consciously adopt this new attitude as a result of "rationally assessing and revising" her values and principles, or "identifying" with it "on the basis of informed, critical reflection, and of intentionally fostering new values and pro-attitudes" in accordance with her considered evaluative judgments (pp. 166-167). According to Mele's theory, her standing capacities for self-control were bypassed and her new attitude was not acquired authentically. If the aversion becomes practically unsheddable, perhaps because of deep physiological causes, then Beth is compelled to have the new value and does not possess it autonomously. Conceivably, it might play a role in a well-informed deliberate decision of Beth's, for instance, a negative vote on a personnel matter. If responsibility requires autonomy then Beth is not responsible for her vote, an odd result given that this manner of attitude-acquisition does not seem unusual. Perhaps many of us have picked up such singular aversions or desires by being unexpectedly exposed to a novel confluence of environmental stimuli. Yet, unless we take the drastic expedient of ruling out "moral luck" altogether, they do not seem to be automatic barriers to responsibility for actions ensuing from them.

Recall that a victim of brainwashing lacks autonomy because its pro-attitudes were acquired in a way that bypassed or "destroyed" a capacity to develop into an individual capable of exercising sort of control over the compelled attitudes (p. 168). Suppose Beth acquired her aversion to people with high-pitched voices by a process of critical evaluation, yet had she been given the right sort of education in primary school then she would have developed in such a way so as to control her newly acquired aversion. So, she had the capacity to develop into a person with such control, but because of the education and upbringing she actually received, the capacity was bypassed or destroyed. Consequently, she is compelled to have that aversion and is not autonomous. This result seems far too sweeping: perhaps at the time of our births, most of us could have

developed a capacity to exert control over a wide variety of pro-attitudes even though we did not. In our early development each of us is subjected to physical and social forces of which we are largely ignorant, over which we have no control, yet from which we acquire values, beliefs, motivations, and capacities for rational evaluation that subsequently guide our choices and actions. These forces "destroyed" any capacity to become a different sort of person with self-control regarding any unsheddable pro-attitude that we happen to have. Consequently, every unsheddable pro-attitude is compelled, and anyone with firm unshakeable principles of action ends up being inauthentic and non-autonomous, e.g., Ann, whose unsheddable values for industriousness "were acquired under her own steam" (p. 155).<sup>6</sup>

To avoid this unhappy result, let us amend Mele's account so that destruction of a capacity that never develops is neither an instance of psychological compulsion nor a barrier to autonomy. The second horn of the dilemma now arises. Mele states that any agent who has ideal self-control, exercises that control, is mentally healthy, and satisfies the compatibilist trio with respect to a given pro-attitude, is strongly psychologically autonomous but not CNC-controlled (p. 189). Why not? Mele allows that a CNC controlled agent--call him "Ned"--can exercise ideal self-control and be mentally healthy (pp. 122-126). Nothing precludes Ned from satisfying conditions 2 and 3 for psychological autonomy. Presumably, he can also satisfy condition 1, for on the amended account, the manipulators might endow him from childhood or early fetal stages with capacities for critical reflection that he regularly exercises in regards to his values and priorities. That his reflections are themselves the product of manipulation does not imply that they lack causal efficacy within his psychology. Indeed, the manipulators might utilize these causal patterns in controlling Ned. But then Ned meets all the conditions for strong psychological autonomy despite being CNC-controlled.

Curiously, Mele is warm to this possibility when he allows that an adult, Fred, might be created with a certain set of sheddable desires and values that his creator knows will eventuate in certain sorts of behavior. Since Fred can reflect intelligently on these sheddable attitudes then he satisfies the three conditions for psychological autonomy, and by exercising ideal self-control and being mentally healthy, is strongly autonomous. Mele contends that compatibilists should accept this result and not conclude that Fred is controlled by his creator:

In one respect, Fred is like any autonomous agent at a deterministic world: his path is causally determined. He is special in having been endowed at the time of his creation with a collection of motivational attitudes for his creators' own purposes. But since these are sheddable attitudes, this detail of his creation does not render him non-autonomous, on the assumption that compatibilism is true (pp. 190-191).

Contrary to what Mele says, even if we grant that Fred has the capacity to examine and revise his values, it remains that the creator has instilled Fred with the values and abilities constitutive of this capacity. Moreover, the creator understands that Fred's pro-attitudes will produce the desired actions as outputs given certain environmental inputs, inputs which the creator knows Fred will be subject to. So, Fred is caused, indeed, deliberately

caused, to behave in a certain way in much the same way that designers of robots program the responses of their machines to various stimuli. In the absence of a precise definition of "X controls Y," this is reason to think that, despite his autonomy, Fred is CNC controlled by his creator.<sup>7</sup>

If this assessment is correct, then a manipulated agent like Fred can autonomously possess pro-attitudes. Is Fred also responsible for what he does? If so, then Mele's account of practical freedom is acceptable only by hard compatibilists. If not, we are back at the initial question of how a manipulated agent differs from a naturally determined agent. Without more being said about psychological compulsion, sheddability, or autonomous action (p. 193) I see no clear answer to this question within Mele's theory. Consequently, it does not provide a decisive rebuttal of the manipulation argument.

## 5. Control and Historicity: Fischer and Ravizza

Incorporating an historically-based notion of autonomy into practical freedom is also a feature of John Fischer's and Mark Ravizza's response to the manipulation argument (Fischer and Ravizza 1998). Unlike Mele, they make autonomy a part of control and their treatment is more closely driven by a concern with responsibility. A responsible agent is one who is an "appropriate candidate for the reactive attitudes," a status achieved on the basis of either behavior or character (p. 6). Persuaded that the Frankfurt examples show that one can be responsible for behavior so long as no "responsibility-undermining" factors are operative in the actual sequence leading up to the action, they advocate a semicompatibilism according to which causal determinism is compatible with moral responsibility but not with a freedom to do otherwise. The control required by responsibility is *guidance control*, understood in terms of properties of the *mechanism* that leads to the relevant behavior, specifically:

- (1) The mechanism must be reasons-responsive: given roughly the same conditions of behavior and with the same mechanism operating, the presentation of various reasons to the agent would have resulted in different behavior. Thus, an agent possesses guidance control over the behavior if there exist scenarios in which the agent would have done otherwise, even though the agent is not able to bring about such scenarios (p. 52).
- (2) The mechanism must be the agent's *own* in that the agent *takes responsibility* for it, that is, he views himself as an agent subject to reactive attitudes by virtue of the way he has behaved as a result of that mechanism (pp. 210-213).

Failure to satisfy (1) can be caused by forms of subliminal advertising, hypnosis, brainwashing, or direct brain manipulation that create physical mechanisms that are not reasons-responsive (pp. 48-49). An agent subject to covert manipulation might satisfy (1), but being unaware of the relevant behavior-producing mechanism it is unable to take responsibility for that mechanism and, therefore, would fail to meet condition (2). If the agent were made aware of it, he would likely not regard it as the causal source of his

behaving as he does and, therefore, would not think himself responsible for what emerges from it.

It is unclear what qualifies as a *mechanism* of behavior on Fischer's and Ravizza's account, much less *the* mechanism of behavior. Presumably not every set of factors that results in behavior B is a mechanism of B, otherwise, causal determinism would rule out (2) from ever being satisfied. Fischer and Ravizza write that "there is no plausibility to the suggestion that all conditions in the past--no matter how remote or irrelevant--must be included as part of the "mechanism that issues in action"" (p. 52). What justifies this claim? Taking 'mechanism' to be synonymous with 'process leading to action' or 'way action comes about' (p. 38), they acknowledge that different mechanisms can operate in a case of any given action, but go on to say that as a "presupposition of the theory" there is "an intuitively natural mechanism that is appropriately selected as the mechanism that issues in action for the purposes of assessing guidance control and moral responsibility" (p. 47). They offer very little by way of saying how this mechanism is "selected" beyond the following. First, the relevant mechanism must not be described in such a way that it entails the action, otherwise it could not satisfy condition (1) of guidance control, that is, the mechanism must be described in a "temporally intrinsic" fashion rather than an "extrinsic" manner inclusive of the action in question (p. 47). Second, since the fact that there are a number of "actions" a person performs at a time does not preclude singling out one of them as relevant in ascribing responsibility, so too, we can pick out which mechanism is the relevant mechanism that "issues in action."<sup>8</sup>

The manipulation argument is supposedly defused by appealing to clause (2) in the characterization of guidance control: a manipulated agent cannot take responsibility for the mechanism that led to action. Here, however, Fischer and Ravizza promise more than they deliver. Observe how they characterize the concept of S's "taking responsibility" for the mechanism that results in his A-ing:

- (1) S sees himself as an agent, viz., "that his choices and actions are efficacious in the world" and, thus, that his own motivational states are the causal source of his A-ing;
- (2) S accepts that his is a fair target of the reactive attitudes as a result of how he exercises agency in A-ing; and
- (3) S's view of himself as satisfying (1) and (2) is based, in an appropriate way, on evidence (pp 210-213).

Now suppose that the behavior of both the CNC controlled agent and the naturally determined agent result from the same sorts of *proximal* mechanisms, that is, from "ordinary practical reasoning" (p. 233) guided, in turn, by their desires, beliefs, reasonings, intentions, and so on. Suppose, further, that each had a hand in forming a good number of these states, even though this formation process in the manipulated agent was engineered by the external manipulators. Since the covertly manipulated agent might take responsibility for the proximal states that led to his action (p. 234), why don't these qualify as the mechanisms of action?

Fischer's and Ravizza's response is that responsible agency requires the satisfaction of certain historical conditions:

In certain cases involving direct manipulation of the brain (and similar influences), it is natural to say that the mechanism leading to the action is not, in an important sense, the agent's own. (p. 230)

We can readily agree that this is so for a mechanism leading to action, namely, the manipulation itself. But why is manipulation the relevant mechanism if another reason-responsive mechanism is operative for which the covertly controlled agent does take responsibility, namely, his own deliberations? What is relevant, they insist, is "how that mechanism has been put in place" (p. 231), and they motivate this by discussing three cases of covert manipulation (pp. 230-236). The first can be dismissed because irresistible desires are implanted that undermine reasons-responsiveness. In the second, while the implanted desires are strong, the mechanism leading to action, namely, the manipulation that induces the desire, is not one for which the agent takes responsibility. Why not? Because the agent does not know about the manipulation and, hence, "has not taken responsibility for the kind of mechanism that actually issues in the action" (p. 233). This response assumes that the relevant mechanism must be the manipulator's machinations but, unfortunately, does not explain why the proximal psychological states are not equally relevant.

The treatment of a third case is more puzzling. Here, the dispositions that constitute "taking responsibility" are themselves implanted. Fischer and Ravizza conclude that condition (3) above might not be satisfied, yet they decline to specify what an "appropriate way" of basing one's beliefs upon evidence is, saying only,

This condition is intended (in part) to imply that an individual who has been electronically induced to have the relevant view of himself (and thus satisfy the first two conditions on taking responsibility) has not formed his view of himself in the appropriate way. (p. 236).

This is no answer at all. Since an appropriately formed view of oneself can be causally determined (p. 236), then unless the appropriate basing relation is characterized in a rigid externalist fashion so as to rule out the presence of total manipulation, it is unclear that there need be any difference between the totally manipulated agent and the naturally determined agent with respect to how they obtained that evidence. The authors are aware of the externalist expedient (pp. 236-237 n.31) but do not develop it. I am skeptical that a difference with respect to evidence gathering and assessment can be secured given that total manipulation can duplicate much of the ordinary causal processes involved in conscious experience, belief acquisition, retention, and evaluation.<sup>9</sup>

If covert manipulation is the mechanism that generated the action then, to be sure, it is not something for which the agent can take responsibility. But it is equally true that agents generally do not take responsibility for the processes of character formation--e.g., parental guidance, public education, peer influence, emotionally-charged delights and

traumas of youthful experience, or the family-building projects of grandparents. Why aren't these antecedents the relevant mechanisms that issue in action? How do such processes differ in any way that is relevant to moral responsibility? These are the questions with which we began. The answer in terms of "taking-responsibility" won't do since agents rarely, if ever, take responsibility for remote processes that caused them to be the sorts of agents they are, yet might take responsibility for their practical reasoning so produced. And it will not do to locate the relevant mechanism within an internal state if external manipulation undermines responsibility. Hence, without specifying what determines the uniqueness of the mechanism that is relevant to responsible action and what an appropriate way of basing one's view of one's own agency upon evidence consists in, I conclude that the Fischer-Ravizza attempt to preclude the covertly manipulated agent from being morally responsible fails. Since it does not obviously preclude the manipulated agent from taking responsibility, then they are absolutely right to admit that "we cannot pretend to have a decisive defense of compatibilism." (p. 236).<sup>10</sup>

## 6. When Historical Considerations Matter

Let us look more closely at how manipulation affects responsibility. Some writers shift easily between talk about an agent's "being responsible" to it being justifiable to "hold" the agent responsible in the same breath (for instance, Mele 1995, pp. 159-161). On the face of it, whether a person can be justifiably held responsible is different from that person's actually being responsible, since those justified in holding S responsible for A-ing might be mistaken about whether S satisfied the responsibility conditions. One might insist that there is no difference here and take responsibility to be constituted by the justification for holding people responsible, but it is more commonly thought that an agent's responsibility is a matter of being worthy of (deserving of, being an appropriate candidate for) praise or blame (Wallace 1994, Copp 1997, Haji 1998, Fischer and Ravizza 1998).

Yet the formula "S is worthy of praise or blame" is itself ambiguous. A good deal has recently been written about the actions of "praising" and "blaming" as responses to behavior that include reactive attitudes or emotions (e.g., gratitude, indignation, resentment, etc.) and, perhaps, sanctioning actions (punishment, reward, verbal praise or criticism, etc.) directed towards agents. These reactive responses, as we may call them, are constitutive of holding an agent accountable (Strawson 1962; Wallace 1994, chp. 3). Accordingly, "S is worthy of praise or blame" can mean that some agent X would be justified by certain standards for either blaming or praising S for what he did or omitted. X must occupy a position of proper authority and possess evidence that S satisfied the responsibility conditions with respect to A, including that his A-ing is subject to those standards, for this to be so (Watson 1996, 235-236). So construed, being worthy of praise or blame is a property that S possesses in relation to agents within a specified normative framework, hence, partly constituted by factors external to S. If the normative framework is one of morality, then we are speaking of moral responsibility from an external perspective or, alternatively, of moral accountability.

Judgments from the external perspective must be distinguished from what Gary Watson has called "aretaic judgments," assessments of an agent's "excellences and faults" that might be true independently of anyone's being justified in any reactive responses towards the agent (Watson 1996, 231). Such judgments consider moral blameworthiness and praiseworthiness as monadic properties determined by the agent's motivational, cognitive, and intentional states, and character traits, so that "S is worthy of praise or blame" means that S is either morally virtuous or vicious in acting as he does. Here it is essential that the agent actually satisfy the responsibility conditions. The dependence upon inner states--so prominent in Kant's treatment of moral agency and the good will--allows us to speak of moral responsibility from an internal, aretaic perspective, that is, from a concern with moral character.<sup>11</sup>

It should not be concluded that there are separate kinds of moral responsibility here; there might be just one phenomenon with different aspects. Indeed, the requirement of mens rea in the legal setting suggests more than an occasional overlap.<sup>12</sup> Yet, justification for responsive reactions requires more than evidence in support of certain aretaic judgments. The practice of holding people responsible also depends upon what can be called the responsive value of particular responsive reactions, that is, their value as mechanisms of correcting, reinforcing, or deterring various types of behavior, attitudes, or character traits. Of particular importance is the responsive value accruing to reactive attitudes and emotions themselves. In plain fact, people want to be the objects of favorable reactive attitudes and not the targets of the negative ones, and will often act accordingly quite apart from the prospect of overt sanctions. Moreover, the reactive attitudes gain value through their ability to move their possessors to act in ways that they might not otherwise do.<sup>13</sup>

To achieve a workable conception of justification, let us speak of a reactive response of type R as being standardly correlated with a responsive value V, that is, with the customarily expected benefits of R-type responses regarding the deterrence, reinforcement, or correction of certain sorts of behavior. This can be relativized further by correlating R-type responses of a population P with V in circumstances of sort C. Accordingly, a general practice of responding in an R-manner to A-type behavior in circumstances or sort C is morally justified within a population P just in case an R-response by members of P to someone's A-ing is standardly correlated with effects having a sufficiently high degree of positive responsive value V. Then, a particular R-type response by X to S's A-ing is morally justifiable when X possesses adequate evidence that (i) the general practice of R-responding to A-type behavior in circumstances of sort C is morally justified; (ii) circumstances or sort C obtain; and (iii) S satisfies the responsibility conditions with respect to A. An additional qualification is possible if we want to make explicit X's office or position of authority with respect to S's A-ing in those circumstances. Also, since our assessments of the responsive values associated with given practices are constantly being updated and refined, then what practices are justifiable responses to which behaviors is subject to variation over time inasmuch as customarily expected benefits change.

Historical factors have an undeniable bearing upon both moral character and moral accountability. Facts about how a person's mental states or character traits were formed and developed are relevant to his or her moral virtue and vice. For instance, one might be responsible for a murder committed today but planned five months ago, and this is important to explaining the differences in the degree of vice involved in that premeditated murder and a case of manslaughter brought about by a sudden provocation and unanticipated homicidal urge. Again, one who spent years overcoming deeply inculcated prejudices in an effort to become a more tolerant person might exhibit greater moral virtue in permitting certain behaviors in his presence than one who was the beneficiary of tolerance-training since youth, even though their intentions and obligations concerning those behaviors might be the same on given occasions. Personal history partly determines the precise degree to which an agent is reprehensible or laudable.

Because these historical factors concern the change, formation, and integration of attitudes and character traits, they are internal to the agent's character and development. By contrast, moral accountability is also attuned to external situations. For example, the mere fact that I intentionally refrained from showing up at a party to which I was invited, but not obligated to attend, is not enough to make me culpable, nor for others to be justified in negative reactive attitudes towards me. But I may be blamed for failing to attend if five months ago I had promised the host to attend. The extent to which I merit a responsive reaction depends not only upon my earlier promise but also that I have not done anything to cancel my agreement in the meantime. Again, suppose an instructor gives a passing grade to a negligent student upon the student's payment of \$500. If this comes to be known by those with the proper authority then standard punitive responses are in order. However, if the teacher were threatened at gunpoint to record a passing grade then the historical circumstance of coercion is reason to override the standard responses. A similar conclusion holds if it was learned that the teacher's recording of the passing grade resulted from involuntarily ingesting a drug which produced irresistible sympathetic desires to accommodate all student requests.

The four responsibility conditions are sensitive to historicity. By failing to attend a party I neglect a present obligation fixed by my past promise, and he who would blame me must be apprised of this promise. If I forgot the promise, a different reaction might be called for, but only because forgetting is a relevant state accompanying an omission. The moral demands upon a coerced teacher can differ from those of the uncoerced teacher given that it is more important to avoid the threatened penalty than to comply with the demand. Coercion is relevant in determining what obligation is binding, thus, in deciding whether the obligation condition is satisfied. In the case of the drug-induced irresistible urge, it is control over one's behavior that is diminished because one is unable to either resist the urge or shed it. Again, knowledge of the agent's history is central in determining the amount of control possessed and what reactive responses are justified. Different responses would be called for were it learned that the teacher was careless in ingesting what he or she did, or deliberately took the drug without knowing what it might produce, or deliberately took the drug knowing what effect it would have.

How does manipulation of psychological states affect responsibility? Everything depends upon what the effects of the manipulation are. Consider moral character first. It is assumed by the manipulation argument that a totally CNC controlled agent can have a coherent psychological profile. He might be endowed with a relatively consistent set of pro-attitudes, possess a considerable amount of self-control, regularly review and reorganize his priorities and commitments in accord with his accumulating experience and reflection, and act in response to well-informed practical reasoning. If so, does he possess a moral character? Suppose a globally manipulated agent, Dan, believes that torture is morally wrong yet takes on a job as an eye-gouger because he enjoys the spectacle of human torment, even though the desire to witness suffering is not irresistible. Does Dan act wickedly? Imagine that he acts with a firm sense of open alternatives, viz., that he can refrain from eye-gouging for all he knows. Is his choice morally depraved? I see no reason to think otherwise. Yes, Dan is the unfortunate victim of his manipulators, his wicked character is engineered, but it is a character, and it is wicked nonetheless. One can be manipulated to wish wickedly just as one can be manipulated to believe falsely. Were Dan manipulated into believing that eye-gouging is a great service to mankind, his character might be less reprehensible, but no less engineered. His situation would be little different from that of an Aztec priest whose commitment to the virtues of human sacrifice was instilled through early indoctrination and reinforced by a steady barrage of praise, honors, and rewards. We shudder equally at the spectacle of either.

One might object that Dan is devoid of any moral character because he is not sufficiently autonomous, his choices are not his own. It was pointed out in section 2, however, that if Dan takes eye-gouging to be "open" for him then he presumes the choice of that profession to be "his own." If it is added that Dan "identifies" with his choice, "takes responsibility" for it, and even possesses "ideal self-control," as it was argued that a totally manipulated agent might, then I claim that he has the autonomy needed to underscore his moral depravity. To paraphrase Kant: one who acts with a sense of openness--implied by the control condition--and with an acceptance of a moral demand--implied by the obligation condition--acts "under the idea of responsibility." And one who acts under the idea of responsibility is, thereby, really responsible in the sense of having a moral character.

Things are more complicated when we shift our focus from moral character to moral accountability. Initially, it might seem that one could be no more justified in blaming Dan for his actions than in chastising a hungry lion for snaring an unwary tourist. However, should Dan's actions and the salient circumstances yield no obvious signs of being subject to covert control, and there is evidence that Dan satisfies the responsibility conditions, then one can have sufficient reason to blame Dan for his actions, indeed, to apply the strongest sanctions standardly allowable to the perpetrators of such incredibly horrendous deeds. Thus, the mere fact of being CNC controlled does not necessarily defeat the claim that Dan is morally accountable, that is, worthy of blame from the external standpoint.

The situation changes if it becomes known that Dan is manipulated, for then it is apparent that there are additional determinants of Dan's actions beyond his practical reasoning,

namely, the goals, intentions, beliefs, and powers of his manipulators. This information alone casts doubt upon whether Dan satisfies any of the responsibility conditions. Perhaps his behavior appears intentional but isn't, or perhaps his induced moral beliefs are vividly opposed to the demands people would normally take him to be subject to, or maybe the manipulators might trump the control Dan would otherwise exercise. For example, they might induce irresistible desires that do not cohere with his values, or diminish the number of apparent options, or induce beliefs that impair his efficacy with respect to actions that otherwise appear open. Again, if the requirement is for dual control, the manipulators might be counterfactual interveners that would prevent Dan from doing otherwise if he chose or even to be able to choose otherwise. As long as there is evidence of that manipulation interferes with satisfaction of the responsibility conditions, then it is appropriate to judge that Dan is neither autonomous nor responsible in each of these cases.

Typically, if there is evidence of manipulation then an agent may be exempted from blame (or praise) despite satisfaction of the responsibility conditions, and it is here that a requirement of alternate possibilities is important even if the agent's actual control is not dual. Thus, to be justified in blaming or praising an agent for an action, one must be able to discern that an opportunity existed for the agent to have refrained if he or she had chosen. Because it is so difficult to determine what a person's motives, perceived options, and intentions actually are, external indicators are vital in appraising whether the agent satisfied the responsibility conditions, particularly the condition of control. If I surmise that S had no opportunity to refrain from A-ing, for example, if I have evidence that S is subject to counterfactual interveners, then, for all I know, perhaps S did chose to refrain but was prevented from so doing. I would then have evidence that S did not satisfy the control condition or the relevant states condition so to justify my blaming him. To increase my confidence that this was not so and that he did satisfy the responsibility conditions, then the following conditional supplies a criterion to be followed in order to establish justification for holding S responsible:

If S is morally accountable for A-ing at t, then at some suitably prior time, S was able to have refrained from A-ing at t.

Analogous criteria are appropriate for omissions and for consequences of actions. So, if one has evidence that an agent is CNC controlled but does not know if the manipulators would not intervene, then this criterion is not satisfied, one lacks evidence that the agent satisfies the control condition and, therefore, one is not justified in holding the agent responsible.<sup>14</sup>

Moreover, by preventing agents from having a suitable degree of control over their inner states or what ensues from them, manipulation threatens to disrupt the standard correlations between specific reactive responses and the expected responsive values. We know this from experience. Just as disciplinary techniques that usually correct unwanted behavior in children, e.g., verbal disapproval, will not work for those whose neurological conditions leave them mentally impaired, so too, they can fail for individuals who have been brainwashed. In severe cases, negative reactive responses are out of order.

Consequently, information that an agent is subject to covert manipulation typically defeats the evidence one has that standard correlations between types of reactive responses and the expected responsive value will hold, and justification for praise or blame is not forthcoming. I say "typically" for the following reason. If there is evidence that the manipulated agent is manipulated so as not to destroy the standard correlations, perhaps because the agent is pre-programmed to respond in an appropriate fashion or because the manipulators would cause the agent to so respond, then there is reason to think that the usual reactive responses will retain their expected responsive value. In such an event, the responses may well be justified.

## 7. Conclusion

I have argued that the presence of manipulation does not automatically defeat the claim that the agent is morally responsible--worthy of praise or blame--from either an internal or an external perspective. Manipulation does matter whenever (i) it prevents an agent from satisfying the responsibility conditions, or (ii) evidence of its presence undermines the justification of standard reactive responses. Here, then, we have an explanation of those instances where a manipulated agent differs from the naturally determined agent with respect to responsibility, underwriting a soft compatibilist denial of the second premise of the manipulation argument. In all other cases, the hard compatibilist rejection of the argument's initial premise is in order.

## NOTES

1. Versions of the consequence argument are advanced in van Inwagen 1975, 1983, Lamb 1977, and Ginet 1983, 1990. Van Inwagen 1983, pp. 183-188 presents a version tailored to responsibility alone. The manipulation argument is set forth in Taylor 1974, pp. 49-51 and, more fully, in Kane 1985, chp. 3 and Kane 1996, chp. 5.

2. See Section 3 of Kant's Grundlegung zur Metaphysik der Sitten. The presumption of openness is central to the account of ability I offer in Kapitan 1996. Similar conditions

are supported in Dennett 1984, pp. 116-118; Strawson 1986, p. 196; Zimmerman 1988, pp. 21-22; Vihvelin 1988, p. 238; Glannon 1995, pp. 267-9; and Bok 1998, chp. 3.

3. Recent research by Antonio Damasio et al (forthcoming in Neuroscience, November 1999) reports that injuries to the prefrontal cortex can alter a person's the capacity to distinguish right from wrong.

4. Mele notes that one can be morally responsible for an action one did not perform autonomously, but not unless one was at some point or other autonomous agent (1995, p. 140).

5. When subject to psychological compulsion of this sort, the agent not only fails to autonomously develop a pro-attitude, but fails to autonomously possess (retain) it since it is unsheddable. Of course, an agent might autonomously develop a pro-attitude without possessing it autonomously, e.g. a drug addict who autonomously develops a desire for heroin but is subsequently unable to shed that desire (Mele 1995, p. 138).

6. On page 168 Mele claims that if an agent is "magically produced" by some devil to have certain pro-attitudes innately then these states are compelled and not possessed authentically. In such a case, no capacity to develop into an agent capable of controlling those pro-attitudes is bypassed, but then any innate unsheddable value is compelled.

7. In correspondence of October 25, 1999, Mele wrote to me that on page 179 of *Autonomous Agents* he took Kane's description of the CNC-controlled agent as one whose choice "necessarily comes out as the controller plans or intends" (Kane 1985, p. 36), to imply that "some attitude that gave rise to the choice was one that the agent was compelled to have." I assume that Kane is talking about causal necessity here, however, and no compatibilist should take causal necessitation of a choice to imply either its compulsion or that of any attitude giving rise to it.

8. See Fischer and Ravizza 1998, p. 47, n.19. One way of specifying which of an agent's many actions are the ones relevant for responsibility are those that satisfy the obligation and intention/foresight conditions, determined in part, by the agent's intentional and doxastic states, including the agent's previous commitments. If this were how "the relevant mechanism" that issues in action is selected then the agent's psychological states are the determining factors, but these provide no means of distinguishing between a normal agent and a manipulated agent.

9. Fischer and Ravizza suggest in note 28 on pp. 234-235, that no coherent self emerges or develops in a totally manipulated agent. Again, they are short on explanation. Why can't the totally manipulated agent be a "coherent self" if he or she possesses enough self-reflection and self-control? What about indoctrinated-from-youth agent? Why should the child who emerges from a controlled environment, beginning in the womb if we like, lack the ability to develop into a self anymore than a child who developed has been determined under other, more normal, influences? On these points, I find excessive hand-waving.

10. Ishtiyaque Haji (1998) attempts to block the manipulation argument in a way that combines elements of both Mele's and Fischer's and Ravizza's account. Haji speaks of *moral appraisability*, viz., being morally blameworthy or morally praiseworthy for the action (1998, p. 8), arguing that it is historically based (p. 123). S is morally appraisable for doing A iff (1) S exercised volition control over A (did A intentionally and holding constant S's motivation for A-ing and S's evaluative scheme, there is a scenario in which S both intends and performs an alternative to A); (2) S believed that doing A had moral value; and (3) S's doing A issues from actional springs that are authentic or are truly S's own. (1998, p. 237). Haji uses condition (3) to block the manipulation argument, where the "actional springs" are constituted by the "evaluative scheme" in terms of which S assesses reasons for action. The scheme is "truly the agent's own," if it is (i) not normative-wise unauthentic, and (ii) appraisability-wise authentic. These two notions are characterized as follows: An evaluative scheme is *not normative-wise authentic* if it involves destruction or repression of an agent's initial normative agency (i.e., it destroys or replaces an "original evaluative scheme" of the agent) (p. 126). An agent's evaluative scheme is *appraisability-wise authentic* if its pro-attitudinal and doxastic elements (i) include all those, if any, that are authenticity demanding; (ii) do not include any that are authenticity destructive; and (iii) have been acquired by modes that are not authenticity subversive (p. 136). In turn, an attitude is *authenticity-destructive* if having it precludes satisfaction of other conditions (epistemic and control ones) for appraisability; *authenticity-demanding* if having it is required for appraisability; and *authenticity subversive* if its instillation is incompatible with appraisability for later behavior (p. 131). The totally manipulated agent is not appraisability-wise authentic since its initial evaluative scheme is induced and, hence, not normative-wise authentic.

There are two problems with Haji's response. First, it seems that the totally manipulated agent's evaluative scheme can be normative-wise authentic inasmuch as it is the agent's initial evaluative scheme. Second, the heavy reliance upon the notion of "appraisability" in Haji's definitions threatens to violate a desideratum for a responsibility condition, namely, that it be decidable independently of assessments of responsibility (or, in this case, of appraisability). Unless we know what the conditions for appraisability are, we cannot say whether the totally manipulated agent is appraisability-wise authentic. A manipulated agent might satisfy the intent/foresight, obligation, control, and relevant states conditions, and have an attitude of seeing himself in control of his actions and as a suitable subject for moral responsibility--attitudes Haji thinks are required to ensure appraisability for subsequent behavior (p. 130). We might think it obvious that the manipulated agent is not "appraisable," but the criteria Haji provides are not sufficiently independent to justify this claim. For these reasons, I find that Haji is unable to preclude the totally manipulated agent from satisfying the authenticity conditions and, given satisfaction of the other conditions, from being morally appraisable.

11. See Watson 1996, p.231. A similar distinction between moral accountability and moral character can be found in Zimmerman 1988, p. 38, and in Fischer and Ravizza 1998, pp. 8-10, n.12. I have made it in Kapitan 1986, p. 248; 1989, p. 36; and 1996, p. 439. The following example highlights the difference. Suppose Hardy intentionally trips Lefty in order to amuse himself by seeing Lefty sprawl in the dust, an act that Hardy

himself believes is of dubious propriety. Anders and his young son Dieter know--and believe it to be generally known--that Lefty is a notorious fugitive from justice who has repeatedly robbed elderly people in their homes. Lefty is at this moment fleeing his pursuers (who include Anders and Dieter) after having just deprived Granny Seelittle of her most prized and valuable heirlooms. Upon seeing rounding a corner, Hardy trips Lefty. Observing this, thinking that Hardy is among those anxious to stop Lefty's escape, and desiring to set an example for his son Dieter, Anders would be morally justified in praising Hardy for his action. From the standpoint of moral accountability, Hardy is praiseworthy for having tripped Lefty relative to Anders. Yet, motivated as he was, Hardy is also *not* worthy of being praised for tripping Lefty; from the standpoint of moral character, he is blameworthy because he exhibited wickedness in tripping Lefty. So, there are at least two different ways in which someone can be morally praiseworthy or blameworthy for having performed a certain act.

12. Fischer and Ravizza note that the two views may be combined into one "mixed" view of MR if one is an appropriate candidate for reactive attitudes just in case one has a "credit" or a "debit" in one's "ledger of life" (pp 9-10), that is, if one possesses the appropriate excellence or fault. Watson also doubts whether the two perspectives on responsibility can be held apart, though he thinks it important to see "that they have distinct sources." (p. 243). See also, Davis 1979, pp. 137-138), and the discussion of *mens rea* in Perkins 1972, chp. 7, and in LaFave and Scott 1986, chp. 3.

13. In Republic 440c-441e, Plato emphasized that the Spirited element ( $\mu$ ) serves as an ally of Reason when activated through attitudes like anger (also, indignation, gratitude, pride), thereby moving the person to act justly.

14. I have discussed principles of alternate possibilities at greater length in Kapitan 1996, pp. 435-441, where versions are stated in terms of a compatibilist notion of ability.

## REFERENCES

Bok, Hilary. 1998. Freedom and Responsibility. Princeton: Princeton University Press.

Blumenfeld, David. 1988. "Freedom and Mind Control." American Philosophical Quarterly 25/3: 215-228.

Christman, John, ed. 1989. The Inner Citadel. (Oxford: Oxford University Press).

Davis, Lawrence. 1979. Theory of Action. Englewood Cliffs, NJ: Prentice-Hall.

- Double, Richard. 1989. "Puppeteers, Hypnotists, and Neurosurgeons." Philosophical Studies 56: 163-173.
- Dworkin, Gerald. 1976. "Autonomy and Behavior Control." Hastings Center Report 6: 23-28.
- Dworkin, Gerald. 1989. "The Concept of Autonomy." In Christman 1989: 54-62.
- Feinberg, Joel. 1989. "Autonomy." In Christman 1989: 27-53.
- Fischer, John and Mark Ravizza. 1998. Responsibility and Control. Cambridge: Cambridge University Press.
- Frankfurt, Harry. 1988. The Importance of What we Care About.
- Ginet, Carl. 1983. "In Defense of Incompatibilism." Philosophical Studies 44: 391-400.
- Ginet, Carl. 1990. On Action. Cambridge: Cambridge University Press.
- Glannon, Walter. 1995. "Responsibility and the Principle of Possible Action." The Journal of Philosophy 92, 261-274.
- Haji, Ishtiyaque. 1998. Moral Appraisability. Oxford: Oxford University Press.
- Kapitan, Tomis. 1986. "Freedom and Moral Choice." Nous XX, 241-60.
- Kapitan, Tomis. 1989. "Doxastic Freedom: A Compatibilist Alternative." American Philosophical Quarterly 26, 31-42.
- Kapitan, Tomis. 1991. "Ability and Cognition: A Defense of Compatibilism." Philosophical Studies 63, 231-243.
- Kapitan, Tomis. 1996. "Modal Principles in the Metaphysics of Free Will." In J. Tomberlin, ed., Philosophical Perspectives 10, Metaphysics. Oxford: Blackwell, 419-445.
- LaFave, Wayne and Austin Scott. 1986. Criminal Law, 2<sup>nd</sup> edition. St. Paul: West Publishing Co.
- Lamb, James. 1977. "On a Proof of Incompatibilism." Philosophical Review 86: 20-35.
- Mele, Alfred. 1995. Autonomous Agents. Oxford: Oxford University Press.

Perkins, Rollin. 1972. Criminal Law and Procedure. Mineola NY: The Foundation Press Inc.

Schoeman, Ferdinand. 1978. "Responsibility and the Problem of Induced Desires." Philosophical Studies 34: 293-301.

Strawson, Galen. 1986. Freedom and Belief. Oxford: Oxford University Press.

Strawson, Peter. 1962. "Freedom and Resentment." Proceedings of the British Academy 48: 187-211.

Taylor, Richard. 1974. Metaphysics. Engelwood Cliffs NJ: Prentice-Hall.

Van Inwagen, Peter. 1983. An Essay On Free Will. Oxford: Oxford University Press..

Vihvelin, Kadri. 1988. "The Modal Argument for Incompatibilism," Philosophical Studies 53,

227-244.

Wallace, R. Jay. 1994. Responsibility and the Moral Sentiments. (Cambridge MA: Harvard University Press).

Waller, Bruce. 1998. The Natural Selection of Autonomy. Albany: SUNY Press.

Watson, Gary. 1996. "Two Faces of Responsibility." Philosophical Topics 24, 2:227-248