Reasons and the Wrongness of Manipulation

Moti Gorin
Rice University

I. Introduction

People are complicated beings, exhibiting an extremely wide range of behaviors that are due to an equally wide variety of causes. Consequently, there are myriad means available to influence this behavior. We make claims, both true and false. We construct good arguments and bad ones. We make different sounds and facial expressions. We clothe, decorate, situate, and move our bodies in seemingly infinite ways. We make use of tools and other sorts of artifacts. We alter our environment and in so doing stimulate our perceptual, cognitive, and emotional faculties. Each of these means of interpersonal influence can be used manipulatively, though none of them is essentially manipulative. The difficulty that motivates this essay lies in distinguishing between manipulative interpersonal influences and non-manipulative influences, and in explaining what it is about manipulation that gives us reason to avoid engaging in it.

I begin with a discussion of the relationship between interpersonal manipulation and deception, harm, the undermining of autonomy, and the bypassing or subversion of the rational capacities. If manipulation necessarily involves one or more of these (putative) wrongs, then it (or they) would provide both a necessary condition in the analysis of the concept of manipulation and an explanation for why we often judge instances of manipulation to be ethically problematic. After briefly describing accounts of manipulation

that involve these wrongful phenomena, I provide counterexamples to each of them. Next, I begin to sketch a general account of interpersonal manipulation, according to which manipulation is a process of interpersonal influence that fails to track reasons. Manipulation's failure to track reasons distinguishes it from ethically benign forms of influence like rightly-motivated rational persuasion. By deliberately influencing others via processes that fail to track reasons, manipulators behave in a morally objectionable manner.

II. Manipulation and Common Wrongs

Manipulation commonly involves ethically suspect behavior such as deceiving, harming, undermining autonomy, or bypassing or subverting the rational capacities. Hence, it is tempting to think there is some necessary connection between manipulation and these other things. There are also theoretical advantages to insisting on a tight link between them, for though deception, harm, autonomy, and the rational capacities remain to varying degrees contested concepts it is at least fairly clear what the major competing normative ethical theories have to say about them. A necessary connection between one or more of these concepts and manipulation would allow for the derivation of conclusions regarding the nature of manipulation from claims about deception, harm, autonomy, or the bypassing or subverting of the rational capacities. The most interesting ethical questions about manipulation would turn out to be questions about other phenomena whose natures have been more frequently discussed and which are better understood. For example, if manipulation always involved deception, then answers to questions about the ethical status of deception would also serve as answers to questions about the ethical status of manipulation. This would leave us with a relatively tidy way to approach questions about

the normative dimension of manipulation.

In the following four sections I examine the relationship between manipulation and deception, manipulation and harm, manipulation and autonomy, and manipulation and the rational capacities. I argue that though manipulation often does involve one or more of these, it does not always do so. An account of manipulation that reduces its normative significance to concerns raised by deception, harm, and threats to autonomy or the rational capacities will fail to capture much that is interesting and important about manipulation. Such an account will therefore remain incomplete. I will begin with a discussion of the relationship between manipulation and deception and then move on to discuss harm, autonomy, and the rational capacities. My strategy will be to motivate accounts of manipulation according to which these wrong-making features are necessary conditions for manipulation and then to provide counterexamples to these accounts. The central conclusion of this section is that manipulation does not essentially involve deception, harm, the undermining of autonomy, or the bypassing or subverting of the rational capacities. None of these can provide a necessary condition in the analysis of interpersonal manipulation.

1. Manipulation and Deception

The first account I will examine pays special attention to the epistemic features of manipulative interactions and in particular to the role of deception in these interactions. On this account, which I will call the Deception-Based View, manipulation always involves some element of deception. A defender of this view can correctly point out that many paradigmatic cases of manipulation involve deception and that deception may enter into a manipulative encounter in more than one way. First and most crudely, a manipulator

may lie, that is, he may state something he knows to be false with the intention that it be believed to be true. Here is one example of this.

> *Not Credible*: Henry wishes to undermine the credibility of his colleague Elizabeth. He lies to her about various matters on which she rightly takes him to be an authority. Later, when Elizabeth is having a conversation with other experts in Henry's field, she relies on the "information" Henry provided her. The specialists, who correctly judge that Elizabeth is advancing false claims, begin to doubt her competence. The experts' judgments that Elizabeth is an unreliable source of information or that she is incompetent, or whatever, are products of Henry's manipulation.

In this case of epistemic manipulation, Henry has manipulated Elizabeth as well as his peers and his method of doing so included the telling of lies as its central component. Less crudely, a manipulator may say something that is true but which he intends will lead his interlocutor to believe something false. Depending on the other beliefs an agent has and on the context of the exchange, the acceptance of a true belief may lead to her acceptance of a false belief. Here is one such case.

> *Synagogue*: David is romantically interested in Susan and so is his friend Jack. David knows Jack is a committed Catholic who prefers to date other Catholics. David knows that Susan, too, is Catholic but he does not wish Jack to know this, as David would like to reduce the amount of competition he might face for Susan's affection. David recently saw Susan entering a synagogue. Though he knows Susan was there only to meet with the rabbi about an upcoming fundraiser for a non-denominational charity, the next time he has lunch with Jack he mentions that he saw Susan at the synagogue. David intends that this will lead Jack to believe that Susan is Jewish and, consequently, that Jack will come to believe that Susan is not a viable romantic option for him.

David states something he believes to be true and he intends that Jack accept the statement as being true. Nevertheless, David intends that Jack's acceptance of a true claim will lead to his holding a false belief and ultimately that this will lead to the behavior David is seeking from Jack. David's behavior is both manipulative and deceptive but it does not involve a lie.

The Deception-Based View of manipulation captures an important feature of manipulation, namely, that it can "prevent [a manipulee] from governing herself with an *accurate understanding* of her situation."[1] In the cases discussed so far, manipulators do this by causing manipulees to have false beliefs whose content extends beyond the intentions of the manipulator, though of course the manipulators also deceive the manipulees about their intentions (otherwise it would not be easy to deceive them about anything else). But manipulators sometimes prevent manipulees from having an accurate understanding of their situation by causing them to have false beliefs or to fail to have salient true beliefs whose content is limited to the ends at which the manipulator's action is aimed and the role the manipulees play in the achievement of those ends. In such cases, the manipulator's intentions are "masked" though the manipulee is not being deceived about anything external to the intentions of the manipulator. Here is such a case.

> Flattery: Carlos approaches his boss Lucinda at the company holiday party and tells her that her recent restructuring of the company's distribution system was altogether brilliant. Though Carlos happens to believe Lucinda's recent performance really was brilliant, he would have told her this even if he believed her efforts displayed rank incompetence. Carlos knows he is telling his boss something she has already heard from many others and which she already believes and he believes that due to his own limited business experience Lucinda probably will not take his opinion to carry much weight as an evaluation of her work. Carlos believes the only value of his expressing his opinion lies in its potentially causing Lucinda to be positively disposed towards him, and he wants badly for her to be so disposed in light of his recent performance review, during which Lucinda expressed serious concerns about Carlos's ability meet the requirements of his job. Carlos is motivated to appear to compliment Lucinda exclusively by the effect he thinks doing so may have on her attitudes toward him.

Carlos does not deceive Lucinda about his opinions of her work but he does act deceptively insofar he wants it to appear to Lucinda that his comment was motivated by his beliefs regarding the features of Lucinda's behavior that really do justify a compliment,

[1] Buss, Sarah, "Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints," *Ethics* 115 (January 2005) p. 226

and not exclusively by his desire to get into her good graces. Carlos must rightly assume that if Lucinda believed that he was merely trying to ingratiate himself to her his action would be unlikely to elicit attitudes that would benefit him. By masking his intentions with respect to Lucinda's attitudes toward him, Carlos attempts to mislead Lucinda about the purpose of his disclosing (what just happens to be) his opinion to Lucinda. The masking of his intentions is necessary for their satisfaction and is a central element in his plan. Carlos is attempting to "prevent [Lucinda] from governing herself with an accurate understanding of her situation" insofar as the success of his plan—i.e., that Lucinda have certain attitudes about him—depends on her misconstruing the purpose of their interaction. Carlos acts deceptively and manipulatively, though the scope of his deception is limited to the content of his intentions.

In all cases of successful deception, the intentions of the deceiver will to some extent remain hidden. In most cases of deception, the masking of the intentions is of derivative, instrumental importance from the point of view of the deceiver, as the more central aim of the deceiver is the acceptance by the deceived of false beliefs about some state of affairs that is independent of the intentions that lie behind the act of deception. But in other cases of deception, the object of the deception just is the content of the deceiver's intentions. The victim of the deception comes to have false beliefs only about what the deceiver is doing in interacting with her. As the case of Carlos and Lucinda illustrates, it is possible for an agent to speak the truth while nevertheless dissembling, as the content of the propositions asserted (e.g., that Lucinda's performance was brilliant) is independent of the content of the intentions that underlie their assertion (e.g., that Lucinda come view Carlos in a more favorable light.)

When one agent interacts with another agent the latter typically will have

expectations about the intentions of the former and the role she (the latter) plays in those intentions. Generally these expectations are not the product of any explicit statement or agreement but are rather assumed to underlie the interaction. For example, in typical cases of communication an agent expects that her communicative partner adheres to certain norms of discourse, for example that she be neither more nor less informative than necessary, that she speaks with the intention to convey what she believes to be true, that she says only what is relevant, and that she is reasonably careful to avoid saying things that may lead to misconceptions or confusion.[2]

I propose to add to this list a Transparency Norm, which requires that an interactive partner not hide her intentions in interacting when these intentions are relevant to the intentions or interests of the person with whom she is interacting. Unlike the truth-telling norm, which is quite general and has application in most (if not all) contexts, the Transparency Norm may have a more limited applicability, the criteria for which will vary with context. For current purposes, I hope only to have shown how deceptive manipulation may involve a particularly nuanced kind of deception, one in which a manipulee is deceived not about the truth value of what the manipulator is claiming but rather about what both manipulator and (as a consequence) manipulee are doing. Indeed, in all cases of deceptive manipulation, whether the content of the deception is limited to the intentions of the manipulator or extends beyond them, a central aim of the manipulator is to deceive the manipulee about the role the latter plays in the plans of the former. Unlike in non-manipulative deception, where the point of the interaction is to cause false beliefs with

[2] These expectations correlate roughly to the four maxims comprising Grice's Cooperative Principle (quantity, quality, relation, and manner). Grice was attempting to provide a theory of meaning in formulating the Cooperative Principle and examining various failures to abide by the Principle. I do not mean to endorse Grice's semantic theory. I appeal to Grice's categories here because they are helpful in articulating the kind of expectations that are generated in a wide range of social interactions. For Grice's discussion of the Cooperative Principle, see his "Logic and Conversation," *Studies in the Way of Words,* Harvard University Press, 1989, pp. 22-40

content extending beyond the intentions of the manipulator, in cases of manipulative deception such beliefs, if they are at all present, are of derivative value to the manipulator, whose central concern is to mask her intentions and the role the manipulee plays in these intentions. The Transparency Norm would rule out deceptive manipulation as well as most standard cases of deception such as lying and is thus more general than a standard truth-telling norm. It is by playing on the expectations of manipulees, expectations generated by adherence to the Transparency Norm, that manipulators prevent manipulees from governing themselves with an accurate understanding of their situation.

In *What We Owe to Each Other*, Thomas Scanlon discusses how our causing others to have expectations about our behavior can generate moral obligations. In this context, he articulates a principle meant to rule out unjustified manipulation. He calls this principle "Principle M" and it requires that (in certain circumstances) agents not hide their (relevant) intentions in interacting with others.

> *Principle M:* In the absence of special justification, it is not permissible for one person, A, in order to get another person, B, to do some act, X (which A wants B to do and which B is morally free to do or not do but would otherwise not do), to lead B to expect that if he or she does X then A will do Y(which B wants but believes that A will otherwise not do), when in fact A has no intention of doing Y if B does X, and A can reasonably foresee that B will suffer significant loss if he or she does X and A does not reciprocate by doing Y.[3]

According to Scanlon, Principle M is a valid moral principle. This is because, "[c]onsidering the matter from the point of view of potential victims of manipulation, there is a strong generic reason to want to be able to direct one's efforts and resources toward aims one has chosen and not to have one's planning co-opted…whenever this suits someone else's purposes."[4] Here Scanlon voices a concern similar to that expressed by

[3] Scanlon, Thomas, *What We Owe to Each Other*, The Belknap Press of Harvard University Press, 1998, p. 298

[4] Ibid. I think it is plausible that Principle M is indeed a valid moral principle. However, the principle is formulated in such a way as to preclude more than one kind of morally questionable behavior, and thus it is not clear that it best accounts for the wrongness of manipulation rather than some other kind of wrong. First, as Scanlon points out, agent A makes it impossible for B to "direct [his or her] efforts and resources toward aims [B] has chosen." Second, A has

Buss when she says that manipulation can "prevent [a manipulee] from governing herself with an *accurate understanding* of her situation."[5] The explanation for Principle M—i.e., that people have strong reasons to want to be able to direct their energies toward aims they have chosen, and that hiding one's intentions when interacting with others can undermine this ability—may capture one ethically troubling element that is sometimes present when one agent manipulates another. The basic idea seems to be that when one's intentions impact the intentions of others, it can be wrong to mislead others about what one's intentions really are. Scanlon goes on to discuss other more general but related principles that he thinks account for the wrongness of promise breaking and lying, and he claims that these principles are generalizations of Principle M.[6] On his view, unjustified manipulation is a special case of lying, and thus Scanlon seems committed to the Deception-Based View of manipulation.

In each of the cases discussed so far, a manipulator deceives a manipulee by making claims (whether true or false). However, a manipulator may avoid making claims and yet use deception to control the behavior of others. For example, advertisers frequently arrange non-propositional visual and auditory stimuli in ways that associate the products they are trying to sell (or the policies they are trying to promote) with the preferences of members of the target demographic, even when there is no rational or causal connection between the stimuli and the products (or policies) with which they are being associated. Many such cases will clearly count as manipulative even if not all of them do. Or, a

---

intentionally sought to gain an advantage at B's expense, as we are told B will suffer significant loss. Third, A has deceived B about A's intentions, the content of which intentions form the basis of B's decision to behave as A wishes. None of these three things form an essential component of the others—they are conceptually independent. One might commit one of these putative wrongs without committing the others.

5 Buss, Sarah, "Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints," *Ethics* 115 (January 2005) p. 226

6 Ibid., pp. 299-322

manipulator may make changes in the environment which are intended to lead to the manipulee's holding false beliefs and behaving on the basis of doing so. Carol Rovane provides a nice example of this kind of manipulation:

> …you are about to leave the house without your umbrella. And…I decide that it would be amusing to get you to take it…I happen to know that you always take your umbrella on days when your housemates take theirs. I also happen to know that there is an umbrella stand near the door which is usually full of umbrellas, except on days when your housemates have taken them. So I remove all of the umbrellas but yours from the stand with the following aim: you will notice that the other umbrellas are gone, you will infer that your housemates have taken their umbrellas, and you will decide to follow suit by taking yours.[7]

Here the manipulator avoids making any claims at all, and yet the manipulation is deceptive.

The cases presented above are representative of a large class of manipulative actions of the sort captured by the Deception-Based View. However, there are counterexamples to the Deception-Based View.

> *Off the Wagon*: Wilson and Adams are up for promotion, though only one of them will get the job. Wilson is a recovering alcoholic and Adams sets out to encourage a relapse, intending this to disqualify Wilson for the promotion. Adams consistently drinks alcohol in front of Wilson, offers her alcoholic beverages, vividly describes to her whatever benefits there are to drinking and to drunkenness, and so on, all the while making no secret of his intentions. During a moment of weakness brought on by a particularly difficult and stressful event Adams takes a drink, which leads to more drinks, missed days at work, and an overall decreased ability to meet the demands of her job. When the time comes to announce who will be promoted, Adams is told by her managers that her recent poor performance has made it impossible for them to give her the new job and that they have selected Wilson for the promotion.

Wilson has manipulated Adams by engaging her compulsion to drink alcohol. And Adams's awareness of Wilson's intentions does not undermine the intuition that this is a genuine case of manipulation. In this case the manipulator does not deceive the manipulee about anything. The manipulator's intentions are known to the manipulee and no false

---

[7] Rovane, Carol, *The Bounds of Agency,* Princeton University Press, 1998, p78

claims are advanced. Therefore, manipulation need not involve any deception. The Deception-Based View is false.

2. Manipulation and Harm

According to the next account of manipulation I will examine—the Harm-Based View—manipulation essentially involves harm, and this is what provides us with a reason to avoid manipulation. The Harm-Based View accounts for the fact that often when we criticize an instance of manipulation one of the features we single out is the harm done to the manipulee, and it accounts for the fact that manipulators often do advance their own interests at the expense of those whom they manipulate. David in *Synagogue* seeks to increase the likelihood of his getting what he wants (a relationship with Susan) by decreasing the likelihood of Jack's getting what he wants (also a relationship with Susan), and in *Off the Wagon* Adams improves his situation by making Wilson significantly worse off. Scanlon's Principle M involves one agent deliberately gaining advantage at the expense of another agent who, as a result of their interaction, would suffer significant loss. Indeed, it might be thought that the motivation for the Deception-Based View is grounded at a deeper level in a concern about harm. Perhaps a defender of the Deception-Based View mistakes the importance of process (deception) with that of a salient consequence (harm) of that process. In any case, an account of manipulation that takes harm to be an essential normatively-relevant feature will capture some cases of manipulation that are left out by the Deception-Based View, e.g., *Off the Wagon.* It will also explain why manipulation often does involve deception, for people who are mistaken about their situation, for example about the consequences of their behavior, are more likely to behave in ways that are detrimental to their own interests.

Typically (but not always) people resort to manipulating others when they believe other methods of influence will fail. Sometimes there simply are no good reasons that can be given to someone to get her to behave in a particular way—not because she is not amenable to reason, but because she *is* amenable to reason and what is being asked of her is contrary to reason. When an agent believes that some possible action of hers will be detrimental to her interests she probably will be strongly disposed to avoid doing that action and if she has sufficient evidence for her belief and is rational there may be no good argument to convince her otherwise. In such cases, it may be necessary for the person seeking control to manipulate the agent into doing whatever it is she wants her to do. As an effective means of directing people to do voluntarily what is not in their best interest, at least according to their own considered judgment (which may or may not be consistent with their judgment at the time of the manipulated act), manipulation often does involve harm to the manipulated agent.

But the Harm-Based View does not stand up to much scrutiny. Perhaps the easiest way to see this is by reflecting on cases of manipulative paternalism. Though it is difficult non-manipulatively to direct people to act in ways that are inconsistent with their own considered judgments regarding their interests, people are prone to acting against their own interests on their own, sometimes consciously, and manipulation can be used to *prevent* them from doing so. The "libertarian paternalist" policies proposed by Sunstein and Thaler are intended to cause people to behave in ways that benefit them and they do so in ways that exploit irrational (or, weaker, non-rational) tendencies.[8] For example, if a cafeteria manager gets people to eat healthy foods by carefully arranging the order in which the food choices are displayed in the cafeteria, it is plausible that he has manipulated his customers

[8] Sunstein, Cass R. and Thaler, Richard, 2003, "Libertarian Paternalism is Not an Oxymoron", *The University of Chicago Law Review,* Vol. 70, No. 4, pp. 1159-1202

to act in ways that benefit them.[9] Here is a more straightforward example.

> *Dementia*: Mildred, who suffers from dementia, appears to have an infection. Her son Nathaniel wants her to go to the hospital but is unable to persuade her to do so by citing the reasons that support her doing so (e.g., that infections left untreated may be life-threatening, that the hospital is the best place to treat the infection, etc.) Nathan knows that his mother would go to the hospital if she were told to do so by his father. The problem is, his father has been dead for a number of years. However, due to her dementia, Mildred often mistakes her son for her husband. Nathaniel waits until his mother calls him by his father's name and then, pretending to be his father, tells her that he would like her to go to the hospital to have her infection treated. She agrees.

This case raises a number of difficult ethical questions. However, it should be clear that Nathaniel has manipulated his mother and also that he neither intended harm nor likely brought any about. Unless we implausibly stipulate that to manipulate someone is *ipso facto* to harm her, the Harm-Based View will be subject to many similar counterexamples.

3. Manipulation and Autonomy

The third view I will examine is the Autonomy-Undermining View of manipulation. According to this account, manipulation essentially involves the undermining of an agent's autonomy. The Autonomy-Undermining View is more difficult to assess than the previous two accounts. Theories of autonomy vary and thus an account of manipulation that makes autonomy-undermining central will need to specify which notion of autonomy is at issue. Broadly speaking, there are two approaches one may take to autonomy. The first is purely "internalist" in that it seeks to locate autonomy in the relations between an agent's propositional attitudes, irrespective of the source of those attitudes or the processes underlying their acquisition and development. The second is "externalist" in that it looks to the sources of an agent's motivational set and the manner in which members of that set

[9] Ibid, p. 1184

were acquired and arranged, i.e., their history. Externalist accounts may themselves differ significantly in how they distinguish between autonomy-conducive histories and autonomy-undermining histories. In this section, I briefly describe internalist and externalist accounts of autonomy and then argue that whichever of these provides the best theory of autonomy, each of them is consistent with manipulation. Manipulation does not entail the undermining of autonomy.

3a. Internalist Theories of Autonomy

On one influential internalist account of autonomy all that matters is the degree of coherence between first- and higher-order propositional attitudes.[10]An autonomy-undermining theory of manipulation that understands autonomy in this way must insist that manipulation enters as an autonomy-undermining technique only between a first-order attitude and a higher-order attitude. To illustrate, suppose an agent has a second-order desire D2 that some first-order desire D1 of his not move him to action. According to the internalist, a manipulator may undermine this agent's autonomy by, say, altering the intensity of D1 so that D1 is now action-causing for the agent. If the agent acts on D1 despite the presence of D2, then the agent has not acted autonomously. Part of the explanation for this is that he was manipulated, since it is the manipulation that led to the misalignment between the relevant attitudes. According to the internalist theory, an action is autonomous when higher- and lower-order attitudes regarding that action cohere in a specific way, and thus for manipulation to be essentially autonomy-undermining is for it to be essentially coherence-undermining.

[10] See, for example, Frankfurt, Harry, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy*, Vol. 68, No. 1, January, 1971

The problem with trying to explain manipulation by reference to in internalist theory of autonomy is that there are cases of manipulation that clearly do not threaten the coherence of the manipulated agent's attitudes. Drawing on the case provided in the last paragraph, a manipulator may leave D1 alone, opting instead to alter D2 so that it coheres with D1. If the agent then acts on D1 he will have done so autonomously according to the internalist. Similarly, a manipulator may alter attitudes on both higher and lower levels so that an autonomous decision not to do X becomes an autonomous decision to do X. For example, as a result of being exposed to subliminal messages an agent who wants to avoid hurting her friend and also wants to want to avoid hurting her friend might form the desire to slap her friend as well as the desire to be the kind of person who desires to slap her friend. If as a result of this she does slap her friend this would constitute a case of manipulation, though according to the internalist account of autonomy the agent acted autonomously. On this picture manipulation cannot be essentially autonomy undermining, for autonomy is preserved despite the manipulation, or even as a result of it.

I do not believe that manipulation necessarily involves the undermining of autonomy. However, in order to vindicate this claim I will need to do more than merely rehearse some of the well-known objections to internalist theories of autonomy. I will need to show how manipulation is consistent with autonomy as the latter is construed by externalist theories as well.

3b. Externalist Theories of Autonomy

Before discussing any particular externalist theories of autonomy, it is important to note an ambiguity about what 'external' is supposed to denote in such theories. On the one hand, there are questions about the sources from which and the processes by which an agent

came to hold the propositional attitudes or, more broadly, to be in the behavior-underlying states in question. On the other hand, there are questions about the agent's attitudes about those processes. I call theories that focus exclusively on the first class of questions *pure externalist* theories. Such theories seek to distinguish between autonomous and non-autonomous behavior (broadly construed to include the acquisition/holding of propositional attitudes, emotional fluctuations, etc.) by reference to the processes that lead up to the states of the agent that underlie the behavior. According to a pure externalist theory of autonomy the truth of autonomy claims can be determined in the absence of any reference to the agent's attitudes about her own states or the processes that lead up to them. The second class of externalist theories, which I label *mixed theories,* hold that in answering the question of whether or not some agent is autonomous with respect to some behavior we must look at the processes that lead to the behavior as well as at the content of the agent's propositional attitudes. With respect to the propositional attitudes, these theories focus in particular on the content that represents the sources and processes that lead to the development or alteration of the agent's behavior-underlying states. According to a mixed theory of autonomy, an agent cannot be autonomous with respect to some bit of behavior if she does not have (*inter alia*) non-negative attitudes about the processes leading up to the states that underlie this behavior. In other words, the agent must approve of the processes.[11] This relation between an agent's attitudes and the processes that lead to her behavior plays the same role in the mixed account that the relation between lower- and higher-order attitudes plays in the internalist account. That is, it is meant to ensure that in order to be autonomous an agent must in some sense authorize the forces that move her. But unlike internalist theories, mixed accounts require that the salient propositional

[11] At this point I understand 'approve' rather weakly as a kind of (actual or perhaps even hypothetical) pro-attitude.

attitudes have as their content the processes that lead to or underlie the relevant behavior.

Accounts of manipulation that appeal to an externalist conception of autonomy are difficult to assess because manipulation is itself a historical (i.e., external) process, one that is often construed as being antithetical to autonomy by definition. In order to defend my claim that the presence of manipulation is at least sometimes consistent with autonomous behavior I will have to pursue one of two approaches. The first is to argue that externalist theories of autonomy fail and so it does not matter that according to these theories manipulation and autonomy are inconsistent. This would leave the internalist theory standing and (as sketched above) autonomy as the internalist construes it is consistent with manipulation. The alternative approach is to show that manipulation does not always threaten autonomy as understood by externalist theories. I will pursue the latter strategy for two reasons. The first is methodological. I do not want the plausibility of my account of manipulation to depend on the truth of a controversial theory of autonomy. Second, I happen to think history does matter when it comes to autonomy. Some of the standard objections to purely internalist theories are decisive in the absence of any appeal to externalist considerations (i.e., historical processes).[12] However, my claim that manipulation is consistent with autonomy does not require that externalist theories are true, but only that, if they are true, they cannot rule out manipulation.[13]

I will take two routes in supporting the claim that manipulation is consistent with externalist conceptions of autonomy. The first will be to provide cases of manipulation in which, intuitively, no one's autonomy is undermined. Next, I will argue in more general terms that the most plausible kind of externalist theory of autonomy cannot exclude

[12] Here I have in mind certain counterexamples to internalism. Mele provides some powerful ones in *Autonomous Agents* (1995).

[13] I thank George Sher for pointing out that my arguments regarding manipulation and autonomy can remain neutral on the question of which theory of autonomy—internalist or externalist—is the best theory.

manipulation.

Here are two cases of manipulation in which no one's autonomy is undermined:

> *Cafeteria*: The manager of a cafeteria wishes to increase his profits. One way to do this is by getting his customers to purchase items with higher profit margins. Suppose people tend to choose the items they encounter earlier, that is, those placed at the front of the food line.[14]Knowing this, the manager places the more profitable items at the front of the line and places the less profitable ones farther down. Consequently, more people begin to choose the profitable items, just as the manager intended. In this case, at least some customers are manipulated into choosing the more profitable items, and yet intuitively no one's autonomy is undermined.
>
> *Lucrative Suicide:* After a long period of philosophical reflection Jacques becomes convinced that in the absence of God life has no meaning. He also firmly believes that if life has no meaning, he has no reason to continue living, for a life without meaning would be for Jacques little more than a stretch of suffering and boredom. But Jacques believes in God and he believes that God's existence lends meaning to life. Thus, he is motivated to continue living his life. James stands to inherit a nice sum of money upon the death of his cousin Jacques. James sets out to convince Jacques that his theism is unfounded with the intention that Jacques's acceptance of this claim will lead to his suicide. James finds the most powerful anti-theistic arguments available and presents them to Jacques who, after a period of reflection, sees the arguments to the end—the very end. James manipulates Jacques into killing himself and yet Jacques does so autonomously.

These cases show that a person's autonomy can remain intact despite the presence of manipulation in the history of the behavior whose autonomy is in question.

There are general arguments to the conclusion that the most plausible theory of autonomy is a mixed theory and that such theories, like internalist theories and pure externalist theories, render autonomy consistent with manipulation. With respect to the first half of this claim, in order to accommodate some strong intuitions about autonomy, intuitions regarding the importance of the agent's attitudes about her own agential capacities, a defender of an externalist account of autonomy cannot appeal to just those processes that underlie the agent's behavior. This is because even if these processes are free from problematic external interference an agent who is alienated from these processes will

[14] Sunstein, Cass and Richard Thaler, "Libertarian Paternalism Is Not an Oxymoron," *The University of Chicago Law Review,* Vol. 70, No. 4, Autumn 2003, p. 1164

lack a critical component of autonomous agency. She will not conceive of herself as an agent acting independently of problematic interferences.

In the absence of the satisfaction of an attitudinal condition, an agent may meet pure externalist conditions for autonomy[15] and yet falsely believe she is being controlled by autonomy-undermining forces. Or she may be free of any problematic external interferences and yet lack a coherent set of attitudes, i.e., she may not identify with her lower-order attitudes. It may be a necessary condition for self-governance that an agent has a conception of herself as self-governing. It is plausible that an agent's attitudes about her own agency partly determine the extent to which she actually is an agent, and thus an analysis of autonomous agency must make some appeal to an agent's representations of and attitudes about her situation. Consequently, a defender of an externalist theory of autonomy is pushed toward a mixed theory, a theory that incorporates some attitudinal condition, such as a condition requiring that an agent approve of the processes that lead to her behavior.

Thus far I have tried to motivate externalism about autonomy and I have sketched some of the reasons why an externalist might be pushed toward a mixed theory of autonomy. It still remains to be argued that mixed theories render manipulation consistent with autonomy. Here I will draw from literature on autonomy and in particular from work that is critical of internalist theories. As already noted, one of the most powerful objections against internalist theories is that higher-and lower-order attitudes can be brought to cohere in any number of ways, some of which are manipulative. My strategy will be to show that the attitudinal condition in mixed theories, that is, the condition requiring that an agent have the right sort of attitudes about the processes leading to her behavior, is vulnerable to

[15] That is to say, the history of how she came to be in the states she is in and to have the attitudes she has may include no external interferences that obviously threaten autonomy (e.g., brainwashing).

the same problem. It will be easier to see this with an example. Take Christman's analysis of autonomy:

> (i) A person P is autonomous relative to some desire D if it is the case that P did not resist the development of D when attending to this process of development, or P would not have resisted that development had P attended to the process;
> (ii) The lack of resistance to the development of D did not take place (or would not have) under the influence of factors that inhibit self-reflection; and
> (iii) The self-reflection involved in condition (i) is (minimally) rational and involves no self-deception.[16]

Whether or not an agent resists the development of some propositional attitude (condition i) is going to be determined (at least partly) by her other propositional attitudes, so a question arises as to whether the agent resisted the development of these attitudes.[17] The same question then arises with respect to the attitudes that determined whether the agent resisted *those* attitudes. And so on. Condition (ii) may be meant to stop the regress but it can do so only with respect to methods that inhibit self-reflection (e.g., brainwashing). Other methods, such as presenting an agent with a circumscribed set of options, presenting those options in one order rather than another, or even creating a context in which an agent is *more* likely to be self-reflective (e.g., as is perhaps the case with Jacques) are not ruled out by the condition specified in (ii). As far as I can tell, there is no way for an account of autonomy that incorporates an attitudinal condition to exclude manipulation. The only way to exclude manipulation is by jettisoning the attitudinal condition and sticking with a pure externalist view. However, as I have already suggested, I do not think pure externalist accounts of manipulation work. (And even if they do work *qua* theories of autonomy, cases like *Cafeteria* and *Lucrative Suicide* suggest that such

[16] Christman, John, "Autonomy and Personal History," *Canadian Journal of Philosophy*, Vol. 21, No. 1, 1991, p. 11

[17] For now I ignore the hypothetical versions of Christman's analysis. First, I am not yet sure how to interpret them. Second, unless the hypothetical consent is the consent of a idealized agent, hypothetical consent seems to reduce to some set of facts about the actual agent that are quite independent of issues of consent. In short, I am generally skeptical about the ability of hypothetical consent to render an agent autonomous.

theories may not be able to rule out manipulation). Therefore, the most plausible competing accounts of autonomy—the internalist account and the mixed externalist account—construe autonomy in manner that renders it consistent with the presence of manipulation.

4. Manipulation and the Rational Capacities

This section is devoted to an examination of the rather plausible claim that when one agent manipulates another, the former bypasses or subverts the rational capacities of the latter.[18] Though its defenders are not always explicit about whether this claim is supposed to establish a necessary condition for manipulation or a sufficient one (or both), the interpretation in which I am most interested understands the claim as providing a necessary condition. Thus, the central question of this section can be formulated like this: does interpersonal manipulation always involve the bypassing or subversion of the manipulated agent's rational capacities?

I would like first to examine the claim that manipulation is a form of influence that *bypasses* an agent's rational capacities. In order to do so it will be helpful to characterize the rational capacities in some way. There are many challenging philosophical questions about rationality and its realization in agents, for example questions about what sorts of psychological states or abilities contribute to making up an agent's rational self. I cannot here address these questions or provide anything like an exhaustive list of the rational capacities. For current purposes I will rely on what should be a relatively uncontroversial characterization of the rational capacities. By 'rational capacities' I will understand those

[18] See, for example, Cave, Eric, "What's Wrong with Motive Manipulation?" *Ethical Theory and Moral Practice,* Vol. 10, No. 2, 2007, p. 138. and Stern, Lawrence, "Freedom, Blame, and Moral Community", *The Journal of Philosophy,* Vol. 71, No. 3 (Feb. 14 1974), p. 74. George Sher suggested to me in conversation that manipulation is the bypassing of the rational capacities.

capacities that enable agents to assess and revise their beliefs in accordance with the basic canons of logic, to evaluate their options against criteria generated by their values and preference sets, to make adjustments to these sets in light of new information, and to act in accordance with their judgments about what they have most reason to do.[19]

The claim that manipulation always involves the bypassing of the rational capacities may be taken to mean that when an agent has been manipulated her behavior is the product of a mode of influence that did not engage these capacities at all. The examples that lend this view its credibility will sometimes involve a manipulator who *directly* causes the manipulee to be in the state the manipulator wishes her to be in.[20] For example, Harry Frankfurt's famous would-be manipulator, Black, has the power to control the causal processes in his victim's nervous system so that he, Black, can immediately determine how his victim chooses to, and indeed does, act.[21] Al Mele has discussed fictional cases of brainwashing where scientists use sophisticated technology to implant new propositional attitudes into a sleeping agent whose behavior they wish to control.[22] Cases of manipulation like those described by Frankfurt and Mele involve processes that bypass the rational capacities of agents.

There are less fantastic ways by which someone may bypass the rational capacities of the person whose behavior he is seeking to influence. Suppose I find it amusing when you tap your foot and, seeking amusement, I play some upbeat music on the stereo, knowing that you tend to tap your foot when you hear such music. Though in normal

---

[19] This conception of the rational capacities is similar to the Cave's conception of the capacities that render an agent "modestly autonomous." See Cave, "What's Wrong with Motive Manipulation?" p. 138 Thus, my arguments regarding the relation between manipulation and the bypassing or subversion of the rational capacities apply to Cave's account of motive manipulation, as he maintains that manipulation is wrong because it violates Modest Autonomy.

[20] "Being in some state" refers here to having certain propositional attitudes, doing certain actions, feeling certain emotions, being in a particular mood, etc.

[21] Frankfurt, Harry, "Alternate Possibilities and Moral Responsibility", *The Journal of Philosophy,* Vol. 66, No. 23, 1969, pp. 835-836

[22] Mele, Alfred, *Autonomous Agents,* Oxford, 1995

circumstances it is not *ir*rational for you to tap your foot in response to upbeat music, your rational capacities will not typically be involved in the process that begins in my playing the music and ends in your foot tapping. Depending on other details of the situation, I may be manipulating you.

In the examples just described, the manipulator influences the behavior of the manipulee and the rational capacities of the manipulee played no mediating role in the process. The rational capacities of the manipulee were entirely bypassed. In other cases, however, the process that begins in the action of the manipulator and ends in a change in the manipulee does involve a stage at which the rational capacities of the manipulee are actively engaged.

Take the following case:

> *Subject Recruitment:* A medical researcher is struggling to recruit a sufficient number of terminally ill research subjects for his drug study. He truthfully tells prospective subjects that unlike the standard drug they are now taking, the new drug he wants them to take often causes some rather serious negative side effects, but also that an earlier preliminary study found the new drug to extend life for twice as long as the standard drug. The researcher here provides information he believes his potential subjects can recognize as relevant to the attainment of their own presumed ends. He wants to appear to be providing them with a powerful reason to join his study and he must assume they are capable of perceiving the information he gives them as just such a reason. He does not bypass the rational capacities of the potential subjects. In fact, he pursues a strategy that crucially depends for its success on these capacities. But suppose the researcher intentionally fails to inform his potential subjects that the new drug will leave them with two months to live instead of the one month they can expect from the standard drug. By choosing to describe the comparative efficacy of the drugs the way he did (i.e., the new drug extends life "twice as long" as the old drug), the researcher omits a relevant fact (i.e., the life expectancy of the drugs in absolute terms) given the increased likelihood of negative side effects associated with the new drug. After all, for many people two months of discomfort will not be preferable, all things considered, to one relatively comfortable month.[23]

The researcher acts manipulatively while pursuing a strategy of influence that engages the rational capacities of the potential research subjects. Clearly, then, manipulation does not

[23] I owe this case to a discussion with Baruch Brody.

necessarily involve the bypassing of the rational capacities, at least when we understand 'bypassing' as a failure to engage.

It is tempting to reply to this case by pointing out that though the researcher did, strictly-speaking, engage the rational capacities of his prospective subjects, ultimately he intended that they make a decision it may have been irrational for them to make had they been better informed about the consequences of their choices. By presenting their options as he did the researcher exploited the potential subjects' tendency in desperate circumstances to rely too heavily on the one positive-sounding piece of information (i.e., the doubling in life expectancy) he believed would move them to act as he wished them to act, and he neglected to provide information that might have swayed them in the opposite direction. On this view, someone who wishes to avoid acting manipulatively must do more than merely engage the rational capacities of the agent she seeks to influence. She must engage these capacities in a way that does not interfere with their proper functioning. The claim here is that manipulation involves the subversion of the rational capacities. I now turn to an examination of this claim.

4a. Subversion as active interference

There is more than one way to construe the subversion of an agent's rational capacities. First, to influence an agent in a way that subverts her rational capacities might be:

> *Active Interference:* to cause a behavior-underlying change in the agent through a process that *actively interferes* with the agent's rational capacities.

A type of interference common to manipulation is the stimulation of psychological states whose presence is incompatible with an agent's ability clearly and accurately to represent and assess her situation or to act consistently with her assessment. Here are three examples

of this, the first taken from the personal sphere, the second from the political, and the third from the commercial.

> *Theater:* I have grudgingly agreed to attend the opening of a play with you. Halfway to the theater, I "engage[] your sublimated compulsive tendency to check the stove" and you turn back towards home. As a result, we miss the play, as I intended.[24] By stimulating a compulsion of yours I interfere with your ability to follow through with a plan on which (let us suppose) you had rationally settled.
>
> *Legislation:* Some elected officials wish to pass legislation because doing so will allow them to tighten their grip on power while enriching their political patrons. They know this particular piece of legislation will be more likely to gain popular support if it is viewed by a fearful and anxious public as a security-enhancing measure. The officials or their representatives make fear-inducing statements through a compliant media before pushing publicly for their bill, which then passes with little public opposition. The psychological states induced by the officials interfere with the ability of citizens to evaluate the rationale for the bill and to assess its full ramifications.
>
> *SUV:* Advertisers hired to help sell sport utility vehicles create and distribute a television commercial depicting a silent office worker trapped in a tiny, drab cubicle. As joyful music begins to play, the worker is suddenly transported into an expansive wilderness where we see him blissfully freewheeling down a mountain in his SUV, accompanied of course by a beautiful woman. By associating the product with an escape to an exciting and carefree existence far from the dull world of the office, the advertisers hope to capitalize on the yearnings of many alienated workers to be rid of the chains that bind them to their desks. No doubt many of these workers will have to spend more hours at their desks in order to afford the SUVs that promise their liberation. Here the preferences of the manipulees are instrumentalized in a way that ultimately undermines the prospects of their satisfaction.

Everyday examples like these lend support to the view that the subversive dimension of manipulation is best understood as some kind of active interference with the proper functioning of the rational capacities. In the absence of such interference the agents described in these cases likely would have made different choices, choices that probably would have cohered better with their rationally assessed preference sets.

Sometimes, though, a manipulator can avoid actively interfering with the proper functioning of the manipulee's rational capacities. For instance, a manipulator may take

---

[24] Cave, Eric, "What's Wrong with Motive Manipulation?" *Ethical Theory and Moral Practice,* 10, 2007, p. 132

advantage of the many cognitive biases that strongly influence our decision-making. In arguing for the permissibility and desirability of exploiting these biases for the benefit of the people who exhibit them, Cass Sunstein and Richard Thaler provide some nice examples that showcase the ways in which agents may manipulate one another without actively interfering with their rational capacities. To take just one, they cite a study showing significant variation in people's willingness to undergo a medical procedure depending on how information about the outcome is framed. When patients are told that 90 percent of those who undergo the procedure survive they are far more likely to consent to it than are those patients who are told that 10 percent do not survive.[25] Now suppose a physician wants to determine the decision of one of her patients. She is aware of the framing effect and intentionally frames the information she provides to her patient in a way that makes it more likely that this patient will choose the physician's favored option. The clinician does not actively interfere with the proper functioning of her patient's rational capacities. And unlike the case of the medical researcher discussed above, the clinician does not leave out any relevant information. Yet, it is fair to say that she manipulates her patient.

4b. A narrow teleological interpretation of 'subversion'

Cases like this, where a manipulated agent's rational capacities are left to function independently of any active interference on the part of a manipulator, suggest that in order to understand manipulation by reference to the subversion of the rational capacities we need a different notion of subversion. Perhaps to influence an agent in a way that subverts

[25] Sunstein and Thaler, "Libertarian Paternalism is Not an Oxymoron", *The University of Chicago Law Review,* Vol. 70, No. 4, (Autumn, 2003), p. 1161. The paper Sunstein and Thaler cite is Donald A. Redelmeier, Paul Rozin, and Daniel Kahneman, *Understanding Patients' Decisions: Cognitive and Emotional Perspective,* 270 JAMA 72, 73, 1993

her rational capacities is:

> *Narrow Purpose Interference:* to cause a behavior-underlying change in the agent via a process that impedes the agent's rational capacities from fulfilling their function.

This conception of subversion would include many cases of active interference with the rational capacities, cases like those involving the stimulation of certain emotions, moods, or compulsions. It would also cover some of the cases in which these capacities are bypassed, cases like those described by Frankfurt and Mele. This is because both the active interference with and the bypassing of the rational capacities often will impede an agent's ability to achieve her ends, which achievement is the central function of the rational capacities, narrowly understood. This notion of subversion will also include some cases where a manipulator merely exploits an independently existing deficiency in the manipulee's rational capacities. The clinician who makes use of the framing effect to manipulate her patient co-opts what appears from the perspective of ideal rationality to be a flaw in the operation of her patient's deliberative faculties.

But this conception of subversion will not work either. First, a manipulator may exploit a cognitive bias in order to get the manipulee to behave in a way that is consistent with the function of her rational capacities. For example, a doctor may exploit the framing effect to get a frantic or otherwise compromised patient to make the choice she would have made were her rational capacities not compromised. Other cases of manipulation do not seem to involve any function-impeding processes. The case of *Lucrative Suicide* described above is such a case. James manipulates Jacques into killing himself, yet he engages Jacques's rational capacities, he does not actively interfere with the functioning of those

capacities, and he does not exploit any cognitive biases or other inherent imperfections in Jacques's capacity to reflect on his situation, assess his beliefs and values, act in light of his assessments, and so on.

4c. A wide teleological interpretation of 'subversion'

Perhaps what the case of Jacques and his conniving cousin shows is not that manipulation does not impede the rational capacities from fulfilling their function, but rather that the function of the rational capacities should be understood more broadly. Thus far I have assumed that the purpose of the rational capacities is just to help agents achieve their ends, given their current attitudes, values, and preferences. This conception of the rational capacities opens the door to cases in which a manipulator appeals to propositional attitudes with problematic content—for example false beliefs—in order to get the agent who holds these attitudes to behave in ways that are internally consistent with the agent's other attitudes and preferences but which are from an objective standpoint unreasonable. Given Jacques's beliefs, desires, values, and so on, his acquisition of the belief that there is no God may have made it rational for him to kill himself. Nevertheless, we may want to say that his suicide was unreasonable. Perhaps Jacques should not have believed that in the absence of God life lacks meaning, or that suicide is the correct response to a meaningless existence. Perhaps he should not have allowed abstract metaphysical arguments to move him to take such drastic action, even if a warrant for such action was the upshot of his rational deliberation.

When James convinces Jacques that there is no God, he provides Jacques with a motivating reason to take his own life—that is, a reason that plays a role in explaining

Jacques's subsequent behavior.[26] What he arguably does not provide, however, is a reason that justifies Jacques's suicide, a reason an appeal to which renders Jacques's action not only consistent with the attitudes he does have, but consistent with the attitudes he ought to have. The thought here is that the function of the rational capacities is best understood at least in part in terms of their linking up with whatever reasons there are, irrespective of whether or not these reasons currently play any role in the agent's deliberation or action. On this view, to influence an agent in a way that subverts her rational capacities is:

> *Wide Purpose Interference*: to cause a behavior-underlying change in the agent via a process that impedes the agent's rational capacities from fulfilling their function, where the function of the rational capacities is (at least in part) to guide an agent towards behavior that is supported by whatever reasons there are, irrespective of whether or not these reasons currently play any role in the agent's belief and preference sets.

In her essay on manipulation in politics, Claudia Mills articulates a view of manipulation that moves in the direction just sketched. According to Mills, manipulation

> in some way purports to be offering good reasons when in fact it does not. A manipulator tries to change another's beliefs and desires by offering her bad reasons, disguised as good, or faulty arguments, disguised as sound—where the manipulator himself knows these to be bad reasons and faulty arguments. A manipulator judges reasons and arguments not by their quality but by their efficacy. A manipulator is interested in reasons not as logical justifiers but as causal levers. For the manipulator, reasons are tools, and a bad reason can work as well as, or better than, a good one.[27]

According to this account, James has manipulated Jacques because he has knowingly disguised a bad reason or faulty argument to commit suicide as a good reason or sound argument to do so. But has James done this? It seems not. Rather than presenting God's non-existence as a good reason for suicide, he exploited Jacques's belief that it was such a reason. Thus, Mills's proposal needs to be amended to say that a manipulator either knowingly offers bad reasons or arguments as good ones or exploits the manipulee's

---

[26] Parfit, Derek, "Reasons and Motivation", *The Aristotelian Society*, 77 (Supplementary Volume): 99-130

[27] Mills, Claudia, "Politics and Manipulation", *Social Theory and Practice,* v21, Spring 1995, pp. 100-101

already mistaking the former for the latter.

There are other problems with Mills's characterization of manipulation. First, not all instances of manipulation involve the offering of reasons or arguments, whether good or bad, sound or unsound. It would be manipulative for someone to smoke frequently in front of a recovering nicotine addict, to leave cigarettes where the addict would be sure to find them, to do things that would make the addict experience stress, and so on, all with the intention that this will lead the addict to start smoking again. But in this case the manipulator does not offer the manipulee any reasons or argument to begin smoking again. Nor does the manipulator instrumentalize the manipulee's antecedent beliefs that there are such reasons or arguments, for we may safely assume that the manipulee quit smoking at least in large part because he lacked such beliefs.

Second, there is a tension between, on the one hand, Mills's observation that manipulators judge reasons and arguments by their causal efficacy and not their justificatory quality and, on the other hand, her central claim that manipulation is a matter of passing bad reasons or arguments off as good ones. She rightly points out that as a causal lever "a bad reason can work as well as, or better than, a good one" but she does not note that the converse of this is true as well. That is, *as a causal lever a good reason can work as well as, or better than, a bad reason.* If a manipulator is indifferent to the justificatory quality of reasons, caring only about their causal efficacy, then it seems that she should use good reasons—that is, reasons that really do justify—when these do a better job at getting her what she wants. When she does this the justificatory quality of the causally effective reason will be merely incidental for her. Yet it does not follow from this that she knowingly disguises a bad reason as a good one.

I will argue below that manipulation is in some sense insensitive to the normative,

as opposed to the causally efficacious, dimension of reasons. This means that when manipulees behave in ways that are supported by reasons it is in some sense a matter of "normative luck." With her emphasis on manipulators' exclusive focus on the causal efficacy of reasons and arguments as opposed to their justificatory quality, Mills seems to be saying much the same thing. But she also seems incorrectly to understand this claim as implying, or being implied by, the independent claim that manipulators disguise bad reasons or arguments as good ones.

If manipulators sometimes traffic in good reasons, then even the wide teleological interpretation of what it is to subvert the rational capacities fails. According to this interpretation, to subvert the rational capacities is to cause a behavior-underlying change in the agent via a process that impedes the agent's rational capacities from fulfilling their function, where the function of the rational capacities is enlarged to include the agent's satisfaction of the demands of the objectively reasonable. A manipulator may engage the rational capacities of a manipulee in a way that neither interferes with those capacities nor exploits one of their existing flaws, and the manipulee may end up believing, desiring, or acting consistently with her own attitudes and preferences as well as with the attitudes and preferences she would have if she were ideally informed and rational.

This last claim, if it is true, means that interpersonal manipulation does *not* always involve the bypassing or subversion of the manipulated agent's rational capacities. One final example is needed to vindicate this claim.

> *Trust Me:* Suppose I intend to tell you a lie two months from now. The lie is going to be so egregious that I am now not very confident that you will believe it when the time comes. In order to gain your trust, over the next two months I offer you sensible advice, I convince you about various matters by constructing sound arguments, I make many true and easily verifiable claims, I criticize others when they lie, and so on.

It is plausible that when I give you a sound argument tomorrow or next week I am manipulating you. If a year from now you read my journal and discover my plot, it will be perfectly reasonable for you to judge that I was manipulating you during these two months, that I manipulated you when I gave you sage advice and good arguments. And yet, while I manipulate you I engage your rational capacities, and the attitudes you take on as a result of my manipulation are the same attitudes you would have if you were ideally informed and rational.

The claim that manipulation essentially involves the bypassing or subversion of the rational capacities appears attractive, at least initially. Manipulators sometimes influence manipulees without engaging their rational capacities at all. Often manipulators actively interfere with the rational capacities or exploit their flaws. And manipulators can make it the case that the agents whose behavior they are influencing believe, desire, or act in ways that are contrary to reason. But if the arguments I have presented thus far are sound, then manipulation need not involve any of these things.

III. Manipulation, Reasons, and Luck

Towards the end of the last section I claimed that sometimes manipulated behavior is supported by good reasons, but that when this is the case it is in some sense a matter of "normative luck." In this section I begin to elaborate on and defend this claim. I believe the plausibility of the previous account—that is, that manipulation involves the bypassing or subversion of an agent's rational capacities—is grounded in our sense that manipulation does not track reasons in the way some other, less problematic forms of interpersonal influence do.

There are at least two ways in which manipulation can fail to track reasons. First, a

manipulator may seek behavior she believes to be, from the perspective of the manipulee, unsupported by reasons. I will refer to behavior the manipulator does not believe to be to be supported by good reasons with regard to the manipulee as *unreasonable behavior* and manipulation that aims at such behavior *unreasonable manipulation*. The most obvious kind of unreasonable manipulation is when a manipulator intends the manipulee to behave in ways the manipulator knows are not supported by good reasons. For example, very often manipulation unjustifiably harms the manipulee or advances the interests of the manipulator at the expense of the manipulee's interests. Scanlon's Principle M is, I believe, intended to rule this kind of manipulation. In such cases there is no good reason for the manipulee to do what the manipulator wants her to do. The manipulator is in no way motivated by reasons that support the behavior because there are no such reasons and the manipulator knows this. Many cases involving the bypassing or subversion of the rational capacities of the manipulee fall into this category. When an agent believes she has good reason to do something and is able and motivated to do it, getting her to fail to do it will often require interference with her deliberative or agential capacities. Unreasonable manipulation fails to track reasons because it is no way part of the manipulator's plan that the manipulee's behavior be supported by reasons.

Things are more complicated when the manipulator believes the behavior he is seeking from the manipulee is supported by good reasons. I will refer to this kind of manipulation as *reasonable manipulation*. There are at least two subcategories of reasonable manipulation. The first includes cases of manipulation where the manipulator is motivated by the reasons that support the behavior of the manipulee. Here the manipulator's end is that the manipulee does what she has reason to do *because* she has reason to do it. I will refer to such manipulation as *paternalistic manipulation*. In cases

like this an explanation of the manipulator's behavior will make reference to the reasons the manipulator believes support the behavior of the manipulee. In other words, the motivating reasons of the manipulator refer to the normative reasons supporting the manipulee's behavior. A doctor who intentionally frames information in a way that increases the probability that her patient will choose a procedure that will further the patient's interests *because* the doctor believes it furthers those interests practices paternalistic manipulation.

The second subcategory of reasonable manipulation is comprised of cases where the manipulator aims to get the manipulee to behave in ways that are indeed supported by reasons, but where this support is not itself something that independently provides a basis for the manipulator's motivation. In these cases of *non-paternalistic reasonable manipulation* the fact that the behavior of the manipulee is supported by reasons either plays no role in the motivations of the manipulator or plays a merely instrumental role. The case of *Trust Me,* in which I provide you with sound arguments and sensible advice as a means of gaining your trust, is an example of non-paternalistic reasonable manipulation, as your behaving in a way that is supported by reasons plays only an incidental role in my intentions, i.e., what I really want is that you come to trust me, not that you come to trust me for good reason.

So, in what sense is it a matter of luck when the behavior of a manipulee is supported by reasons? It is a matter of luck insofar as the process that led to the behavior is not a reason-tracking process. The failure of reasonable manipulation to track reasons is attributable to one of two things, depending on which subcategory of reasonable manipulation is at issue. With respect to non-paternalistic reasonable manipulation, the failure of the process to track reasons can be traced entirely to the motivations of the

manipulator. In these cases the reasonableness of the manipulee's behavior is not an end at which the manipulator is aiming. The fact that the behavior is supported by reasons does not play an independent role in the plans of the manipulator or in the actions structured by these plans. At most, the reasonableness of the behavior motivates the action of the manipulator indirectly, as when the reasonableness is instrumentally valuable with respect to the manipulator's ends. For example, in *Trust Me* I might want you to trust me for the right reasons—i.e., because I provide you with sensible advice and good arguments—but only because I worry that if you were to come to trust me for bad reasons other people might dissuade you from trusting me.

This point can be illuminated by contrasting reasonable manipulation with rational persuasion, a method of interpersonal influence commonly viewed (perhaps especially by philosophers) as realizing some ideal of human interaction. Both methods can be used to change an agent's behavior, and though there are several important differences between them I think one of these differences is especially deserving of attention. Typically, when we set out rationally to persuade someone what we do is motivated not only by its relation to our end narrowly conceived—i.e., that our interlocutor come to have some attitude or to do some action on a particular occasion—but also by the independent value we believe is realized by our means of influence. We believe it is of value that our interlocutor accepts the premises we offer because doing so will lead her to accept a particular conclusion, but we also believe that her acceptance of these premises along with her making certain inferences has some independent value. We believe either that it is intrinsically valuable to have true beliefs and to reason well or that doing so is instrumentally valuable with respect to ends other than (and in addition to) the end at which we are now directly aiming. If in the middle of my attempt rationally to persuade you I suddenly realize that my argument is

unsound I (typically) will stop what I am doing, even if you have not noticed my mistake and are happily going along with the argument. I will stop because I do not just want you to be convinced of what I am saying. I want you to be convinced of what I am saying for the right reasons. But when I manipulate you my behavior is motivated only by the value I believe is realized by the narrow end at which my action is directly aiming.

In other words, in typical cases of rational persuasion the causal efficacy of the means is not a sufficient condition for their use. Here the value realized by the interpersonal interaction derives from more than one source and the agent doing the influencing is motivated by a relatively broad range of considerations, including the value of the means (e.g., true premises, truth-preserving logical inferences) that is realized by or accrues to the agent she is seeking to influence. This value is independent of the value of the influencer's immediate end, that is, that the influenced agent comes to have the particular attitude or to do the particular action at which the interaction is aimed. In cases of non-paternalistic reasonable manipulation the manipulator is motivated by only the causal efficacy of her chosen means of influence as they are related to her immediate end, i.e., that the manipulee behave in some particular way. Again, in the case of *Trust Me,* in which I give you sound arguments and sensible advice, what I care about is that you come to trust me. If I knew in advance that when the time comes you will believe me in any case (e.g., I know you to be extremely gullible) or if I knew in advance that when the time comes you will not believe me in any case (e.g., I know you to be extremely incredulous) I would no longer be motivated to do those things, that is, to give you sensible advice and sound arguments. Sensible advice and sounds arguments are valuable along several dimensions, but the only value in which a manipulator is interested is the value realized by their causal efficacy.

As I argued above, Mills's claim that manipulators judge arguments and reasons "not by their quality but by their efficacy" is overly cognitive, as manipulators do not always give reasons or arguments. However, we can broaden her claim to include other forms of influence. The crucial point is that in cases of non-paternalistic reasonable manipulation, when manipulators choose some method of influence their choice is exclusively (or at least disproportionately) motivated by the narrow instrumental value of that method. A fundamental feature of manipulation, one that is reflected in every sense of the term and in its etymology, is its instrumentality, its connection to something's being handled, used. In the interpersonal context, too, manipulation is related to the notion of using something as a means to some end.[28] I believe the limited range of considerations that motivate both kinds of non-paternalistic manipulation (i.e., unreasonable manipulation and non-paternalistic reasonable manipulation) explains this connection. Non-paternalistic manipulative actions are motivated by a narrow range of considerations and are consequently insensitive to other sources of value that may be realized by interpersonal interactions. Non-paternalistic manipulators care exclusively about making certain changes in the world, and not about whether those changes are supported by reasons from the perspective of the people who realize those changes—that is, the people whose behavior the manipulators are influencing.

The structure of paternalistic manipulation differs from that of non-paternalistic reasonable manipulation. Whereas non-paternalistic manipulators are not motivated by the independent value of the reason-supportedness of the behavior they are seeking to bring about, paternalistic manipulators are so motivated. Their aiming to get the manipulee to

[28] One possibility is that in the interpersonal context manipulation involves the use of agents, of persons, as means. I will discuss this possibility elsewhere. For now, I will leave it open whether an insensitivity to the wider value of the means of interpersonal influence should be identified with a failure to treat persons as ends in themselves.

behave in ways that are supported by reasons and the manipulee's prior unwillingness or inability to behave in this way are what jointly make paternalist's actions paternalistic. Thus, there is a sense in which paternalistic manipulation is a reason-tracking process in a way non-paternalistic reasonable manipulation is not. If a paternalistic manipulator comes to believe that her chosen method of influence will not lead to behavior that is supported by reasons she will either adjust those means in order to ensure that they do lead to such behavior or she will give up her efforts at influencing. The capacity of the means of influence to bring about reason supported behavior is a necessary condition for their being chosen, and thus the influencer will discard these means if she comes to believe they will not lead to reason-supported behavior. Unlike a non-paternalistic manipulator, she will discard these means even if she still believes that they would be causally effective in bringing about the behavior. It is the reason supportedness of the behavior that, for paternalistic manipulators, provides an independent basis of motivation for resorting to the means that lead to the behavior. Unlike non-paternalistic manipulators, paternalistic manipulators are motivated to bring about the mental states or actions they believe the manipulee has reason to have or to do just in virtue of their believing that the manipulee has these reasons (perhaps in addition being motivated by other considerations as well). That there are normative reasons for the manipulee to behave in some way itself provides paternalistic manipulators with a motivating reason to bring that behavior about.

Thus, paternalistic manipulation tracks reasons in a way unreasonable manipulation and non-paternalistic reasonable manipulation do not. Nevertheless, it does not do so the way rational persuasion or other non-manipulative methods of influence do. The difference between paternalistic manipulation and rational persuasion lies in the antecedent inability or unwillingness of the recipient of the influence to behave in the ways the influencer

intends that she behave despite there being reason for her to so behave. When the object of influence is unable or unwilling to recognize the reasons that support some behavior or when she recognizes these reasons but is not sufficiently motivated to act in accordance with them, rational persuasion will be an ineffectual means of bringing that behavior about. Here an influencer who would otherwise choose means that realize a broader range of values (e.g., the value of accepting true premises and reasoning well) may resort to means with only instrumental value. In this regard, paternalistic manipulation is similar to the other forms of manipulation in that the manipulator's choice of means is determined exclusively by their instrumental value in bringing about the behavior. But it differs from these other forms insofar as it is guided by the reason-supportedness of the behavior at which it is aiming.

Sometimes manipulators are motivated by the reason-supportedness of the behavior they seek to bring about. When the reason-supportedness of the behavior does not provide an independent basis of motivation, the manipulation is not guided by reasons in the way ethically benign forms of influence are. When the reason-supportedness of the manipulation does provide an independent basis of motivation, the manipulation is paternalistic. Paternalistic motivation tracks reasons in a way non-paternalistic manipulation does not, but because here the manipulator values the means of influence only insofar as they are causally efficacious in bringing about the behavior, the process does not track the reasons that independently support the means of influence.

The claim, scrutinized earlier, that manipulation necessarily bypasses or subverts the rational capacities, is false. It is possible to manipulate someone while both actively engaging her rational capacities and causing her to behave in ways that are supported by reasons. However, because manipulation does not track reasons in the way non-

manipulative methods of influence do, there is an interesting and ethically salient connection between manipulation and reason. Sometimes manipulation does not aim at behavior that is supported by reasons. Other times, it does so but only incidentally. And when manipulation does aim at reason-supported behavior *because* it is reason-supported (i.e., in cases of paternalistic manipulation), the means are chosen exclusively for their narrow instrumental value. I believe that these relations between manipulation and reasons —relations which form at least a necessary condition for manipulation—are what explain the temptation to think that manipulation always bypasses or subverts the rational capacities.

The most salient elements of an interpersonal encounter in which one agent seeks to influence another are the motivations of the influencer, the particular means of influence chosen by the influencer, and the behavioral states of the agent being influenced. In an ideal kind of interpersonal influence like rational persuasion with the right intentions everything links up nicely: the motivations of the influencer are grounded in the reasons that really do support the behavior she seeks from the person being influenced, the means of influence (e.g., sound argument) reliably "aim at" or "link up with" these reasons, and the states of the person being influenced (e.g., her propositional attitudes) also link up with the reasons that support the behavior. In cases of manipulation, there are breakdowns in these relations, breakdowns that can occur in more than one place. The location of the breakdown will determine whether the manipulation is reasonable or unreasonable, paternalistic or non-paternalistic.

I believe the various wrongs that manipulation often involves—e.g., harm, the undermining of autonomy, deception, and the bypassing or subversion of the rational capacities—are particular manifestations of the more general phenomenon of failing to

track reasons. As I noted above, cases of manipulation in which the interests of the manipulee are set back are examples of unreasonable manipulation, as here the manipulee has no reason to behave in the way the manipulator intends she behaves. In these cases a manipulator will often need to undermine an agent's autonomy or to bypass or subvert her rational capacities, as otherwise the manipulee will be unlikely to behave in the way the manipulator intends her to behave. Manipulators aiming at unreasonable behavior may also rely on deception, as having an accurate understanding of her situation will reduce the probability that a manipulee will behave unreasonably. Thus, unreasonable manipulation may involve any combination of deception, harm, the undermining of autonomy, or the bypassing or subversion of the rational capacities.

Paternalistic manipulation, where a manipulator is motivated by the reasons she believes really do support the manipulee's behavior, may also involve some of the common wrong-making features discussed earlier. A paternalistic manipulator will choose means with only narrow instrumental value—that is, means that are causally efficacious at bringing about the desired behavior but which bear no normative relation to the reasons that support the behavior—because she judges that means with wider value will not be as effective. So, for example, a paternalistic manipulator may make false or misleading claims if she believes that the manipulee's acceptance of these claims will lead to reason-supported behavior. She may stimulate the compulsions of the manipulee. Or she may evoke emotions that lead to the desired behavior despite lacking the right kind of "fit" (e.g., a manipulator evokes sadness from a manipulee because the latter tends to behave reasonably when she is sad). In some cases of emotional manipulation like this, the manipulee's rational capacities are bypassed or subverted because the presence of the emotional state rather than the recognition of the reason that speaks in favor of the

behavior is what plays the dominant role in determining the behavior. In other cases, however, the emotional state is what makes recognition of the reasons possible, and thus functions as a part of the agent's rational capacities. In these latter cases, the rational capacities are not bypassed or subverted, though other wrong-making features may be present (e.g., deception, the masking of relevant intentions, etc.)

A non-paternalistic manipulator engaging in reasonable manipulation will aim at reasonable behavior but here the reasonableness of the behavior is incidental to her intentions. As already noted, non-paternalistic (reasonable) manipulators may aim at reasonable behavior but not *because* it is reasonable, but rather exclusively for some other reason (e.g., that it satisfies some desire of the manipulator, say). It just happens to be the case that the behavior being sought is supported by reasons and it just happens to be the case that an appeal to these reasons is causally efficacious in bringing about the desired behavior. Such manipulation typically will not be harmful. It may, however, involve wrong-making features just as paternalistic manipulation does. Non-paternalistic reasonable manipulation may, for example, involve deception, the bypassing or subversion of the rational capacities, or the undermining of autonomy. In the case where I offer you sound arguments and sage advice only because I wish to gain your trust I mask my intentions, as it is fair to assume that you incorrectly interpret my behavior as being motivated by my recognition of the reasons that support the claims I advance and the advice that I offer.

Here again are the three types of manipulation, distinguished according to the way each of them fails to track reasons:

> *Unreasonable Manipulation*: a process of interpersonal influence in which the influencing agent is not motivated by reasons that, with respect to the influenced

> agent, support the behavior of the influenced agent because the influencing agent does not believe there are any such reasons.
>
> *Non-paternalistic Reasonable Manipulation*: a process of interpersonal influence in which the influencing agent is not motivated by reasons that, with respect to the influenced agent, support the behavior of the influenced agent, which reasons the influencer believes do exist.
>
> *Paternalistic Reasonable Manipulation*: a process of interpersonal influence in which the influencing agent is motivated by reasons that, with respect to the influenced agent, support the behavior of the influenced agent, but where the means of influence bear no normative relation to the reasons supporting the behavior and are chosen exclusively for their ability to cause the behavior at which the influencer is aiming.

IV: The Wrongness of Manipulation

I believe manipulation's failure to track reasons is what distinguishes it from paradigmatically innocuous forms of interpersonal influence like rightly-motivated rational persuasion, and I think this defining characteristic of manipulation is also what renders it morally suspect. If I am right, then some version of what I will call the Manipulation Principle is true.

> Manipulation Principle 1: it is morally impermissible intentionally to influence another agent via a process of influence that fails to track reasons.

This principle is merely preliminary because it does not distinguish between justified and unjustified cases of manipulation. There may be cases where there is nothing even *prima facie* wrong with manipulation. For example, there may be nothing objectionable, *prima facie* or otherwise, about manipulating a severely intoxicated or otherwise deranged gun-wielding person in order to get him to surrender his weapon. Thus, the final principle will have to include a clause that distinguishes between manipulation that is ethically problematic and manipulation that is not. One possibility is that we distinguish between justified manipulation and unjustified manipulation by appealing to the degree to which the mechanisms on which the manipulee behaves are responsive to reasons. On this account, if

there are no reasons-responsive mechanisms to which an influencer might appeal, or if these mechanisms are only very weakly responsive to reasons, then manipulating the agent whose mechanisms these are may be morally permissible.

> Manipulation Principle 2: it is morally impermissible intentionally to influence another agent via a process of influence that fails to track reasons, unless the person being influenced is relevantly unresponsive to reasons.

Manipulation Principle 2 distinguishes between agents who are responsive to reasons and those who are not. Because it seems too much to require that agents limit themselves to reasons-tracking processes of influence when the targets of the influence are insensitive to reasons, this principle allows us to avoid the controversial conclusion that manipulation is always wrong (or even always *prima facie* wrong). However, the principle still problematically allows for the *unreasonable* manipulation of people who are not responsive to reasons. Clearly, the fact that someone is not responsive to reasons does not justify leading her to behave in ways that are not supported by reasons, even if she cannot recognize or react to those reasons. For example, it is impermissible for a doctor to manipulate a patient who is cognitively impaired into taking part in an extremely risky medical procedure that would, at best, provide only negligible benefits to the patient. This would be impermissible irrespective of the extent to which the patient is reasons-responsive. This suggests that the Manipulation Principle needs to be amended in order to distinguish between reason-supported behavior and behavior that is not reason-supported.

> Manipulation Principle 3: it is morally impermissible intentionally to influence another agent via a process of influence that fails to track reasons, unless the person being influenced is relevantly unresponsive to reasons *and* the behavior being sought is, with respect to the person whose behavior will be influenced, supported by good reasons.

Manipulation Principle 3 rules out unreasonable manipulation but it does not rule

out either paternalistic or non-paternalistic reasonable manipulation so long as the manipulee is not relevantly reasons-responsive. Thus, as it stands, the principle is too permissive, as it allows for the reasonable manipulation of anyone who is not relevantly responsive to reasons. There may be cases in which it is impermissible to manipulate a non-reasons-responsive person even though the behavior at which the manipulation aims is supported by reasons. For example, suppose I want you to lend me money and that you have good reason to do so. I can manipulate you into giving me the loan today while you are severely depressed—by stimulating some compulsion of yours, say— and unresponsive to reasons, or I can wait until tomorrow to present you with the reasons that support your giving me the loan. In either case, you will lend me the money. It is plausible that I ought, morally, to wait until tomorrow and that I have acted impermissibly if I manipulate you today.

Manipulation Principle 3 also suggests that the reasonable manipulation of a reasons-responsive agent is always impermissible. This, however, is too strong, as there may be situations in which it is all-things-considered permissible to manipulate a reasons-responsive agent. For example, suppose again that I want you to lend me money and that you have good reason to do so. But this time, I need the money immediately—it cannot wait a day or even another few minutes. I know that you would be happy to loan me the money and that you would do so as a result of your recognizing the reasons that speak in favor of your doing so. However, it would take too long for me to describe these reasons now. Manipulating you into giving me the money—again, perhaps by stimulating some compulsion of yours—would be much faster. In this case, it is at least plausible that my manipulating you is not impermissible, given the extenuating circumstances. This example and the one that preceded it suggest that further refinement to the principle is necessary.

> Manipulation Principle 4: it is *pro tanto* morally impermissible intentionally to influence another agent via a process of influence that fails to track reasons, unless the person being influenced is relevantly *and decisively* unresponsive to reasons *and* the behavior being sought is, from the perspective of the person whose behavior will be influenced, supported by good reasons.

The addition of the "*pro tanto*" clause to the Manipulation Principle makes room for considerations that may override the wrongness of acts that involve manipulation while still acknowledging that there remains (non-decisive) reason to avoid manipulation even in those cases. And by requiring that the target of reasonable manipulation be not only relevantly reasons-unresponsive but also *decisively* reasons-unresponsive, Manipulation Principle 4 can rule out those cases of reasonable manipulation that are impermissible (or *pro tanto* impermissible) despite their aiming at reasonable behavior from reasons-unresponsive agents.

V: Conclusion

Clearly, more work needs to be done. I have not provided a general account of what it is for a process to be reasons-tracking. I have not said anything about how to distinguish relevant reasons-responsiveness from irrelevant reasons-responsiveness, or for what it is for an agent to be "decisively" unresponsive to reasons. And my claim that manipulation's failure to track reasons is what renders it ethically suspect has mainly been supported thus far by my roughly contrasting non-reasons-tracking processes with a morally benign form of influence—rational persuasion—that does track reasons. But of course, it is possible that manipulation differs from rational persuasion in the way I have suggested—that is, the latter tracks reasons while the former does not—but that, nonetheless, this feature of manipulation is not what renders it morally wrong. In order properly to vindicate Manipulation Principle 4, it would be necessary to say much more about the role of reasons

in structuring our interpersonal interactions and also about why it matters morally that we choose reason-tracking processes when we wish to influence others. But for now, I hope to have accomplished only two relatively modest things. First, I have tried to show why some accounts of manipulation that initially might appear attractive—those that appeal to phenomena such as deception, harm, autonomy, and the rational capacities—do not work. Second, I have tried to sketched an alternative theory, one that begins to provide a general explanation of the nature and ethical significance of interpersonal manipulation.