

RICE UNIVERSITY

The Nature and Ethical Significance of Manipulation


by

Moti Gorin

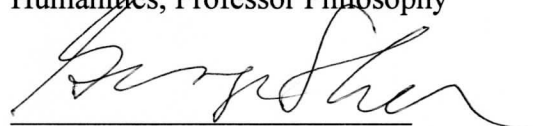
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

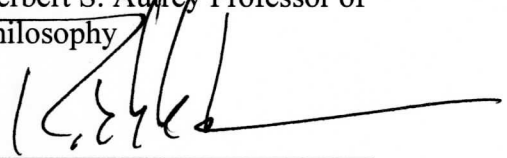
APPROVED, THESIS COMMITTEE:



Baruch Brody,
Andrew W. Mellon Professor in
Humanities, Professor Philosophy



George Sher,
Herbert S. Autrey Professor of
Philosophy



Rick K. Wilson
Herbert S. Autrey Chair of
Political Science

HOUSTON, TEXAS
JUNE 2013

ABSTRACT

The Nature and Ethical Significance of Manipulation

by

Moti Gorin

What distinguishes manipulative interpersonal influence from non-manipulative influence? When is it wrong to manipulate a person and what makes it wrong? I articulate a novel account according to which interpersonal manipulation is a process of influence that deliberately fails to track reasons. To manipulate a reasons-responsive person is to render her detached from an important aspect of reality, namely, her reasons or the considerations that ought to govern her behavior. This is what makes manipulation *pro tanto* morally impermissible (when it is). My account of manipulation provides a helpful framework for thinking through some of the philosophically neglected ethical issues arising out the application of social scientific research on human decision making in the domains of health care and public policy.

Acknowledgements

In writing a dissertation one spends a great deal of time alone, sitting and typing. In this respect it is a solitary activity. The putting of words to paper, though, is really just the final flourish in a complex process to which many people contribute, wittingly or unwittingly. The same probably is true of every intentional and valuable human activity (and some not-so-valuable activities). This point may be trivial but the contributions are not.

The Philosophy Department at Rice University provided me with six years of funding, including travel grants to attend conferences where I presented sections of this dissertation. I am very grateful to the Department for its support and in particular for granting me the Robert Alan and Kathryn Dunlevie Hayes Fellowship, which helped me enormously during my last very busy year. Richard Grandy was instrumental in securing critical funding for me more than once. Thank you, REG. Nico Orlandi and Gwen Bradford helped me navigate the job market and also provided me with good questions and suggestions regarding my work. During my dissertation proposal defense Casey O'Callaghan raised a serious objection that forced me to rethink my view of manipulation. As a result I am sure the account set out below is more plausible than it otherwise would have been. Our department Coordinator, Minranda Robinson-Davis, helped me over and over and over again with many things, not least of which the many logistical steps required of one who is applying to dozens of jobs. Without her help I would not have had as much time to work on my dissertation. No matter how many times I thank Minranda I feel it is never enough.

There have been advantages and disadvantages to living in Austin rather than in Houston during my last couple years at Rice. Certainly one of the biggest disadvantages was that I could not meet in person with Baruch Brody, my primary advisor, nearly as often as I would have liked. Though we did not meet in person often during the period when I did most of my writing, Baruch read my drafts and provided me with a steady stream of incisive criticisms and constructive comments via email or telephone, and he was always ready to meet in person on those occasions when I was in town. Baruch was the first person to draw my attention to the puzzle of manipulation. It was in his seminar that I first began seriously to consider devoting my energies to this important and challenging topic. Baruch has also provided invaluable support as I begin to navigate the professorial waters. I have little doubt that whatever success I have enjoyed thus far in this regard is due in no small part to his efforts on my behalf.

Before I knew what topic I would address in my dissertation I knew that I wanted to work with George Sher. I have spent more time talking philosophy with George than with any other professor at Rice. He is a model of fair-mindedness, a characteristic that is most pronounced—and most valuable—in those cases where you disagree with him. George and I have had plenty of philosophical disagreements over the years and I have always been impressed with his ability to help me develop arguments for conclusions he does not like. In all our conversations he always took me seriously and interpreted my claims charitably. This makes me think George might be an egalitarian, and not the inegalitarian kind, either (no offense, George). I learned much in his seminars and hope one day to write half as well as he does.

Despite not being on my committee and indeed not even teaching at Rice, Mark LeBar, my teacher at Ohio University, gave me very helpful comments on sections of my dissertation. More generally, Mark has for years consistently pushed me toward greater analytical precision and clarity. I always enjoy my discussions with Mark, as he loves to talk philosophy at least as much as I do.

As an undergraduate student at Bradley University I was lucky to take courses with Michael Greene and Daniel Getz. Dr. Greene's passion in the classroom first inspired me to major in philosophy. Dr. Getz's door was always open and he was never short of encouraging words. Had I not had these rich and confidence-building relationships with my professors as an undergraduate student I may not have decided, years later, to enter graduate school.

I am grateful to Professor Rick Wilson for serving as my outside reader. Despite the long periods during which I sent him no new material, Professor Wilson always quickly responded to emails to express his willingness to help. I regret that we did not have more time to discuss manipulation in the political context.

It is practically a cliché to say that as a graduate student one learns more from one's graduate student peers than from one's professors. As an aspiring professor I find this claim a bit troubling and as a graduate student I find it a bit false. I have been fortunate to have had both great teachers and great peers, such that I cannot seem easily to establish who taught me what. In any case, I have benefited greatly and along many dimensions from my relationships with my graduate student friends. Here is a list, in no particular order, of just a few of the graduate students (or former graduate students) who in one way or another helped me complete this dissertation while having fun along the

way: Stan Husi, Jacob Mills, Jennifer Bulcock, Anthony Carreras, Jeremy Garrett, Derrick Gray, Dani Wenner, Joe Adams, Dustin Tune, Jesse Slavens, Justin Ho, and Phil Robichaud.

I presented sections of Chapter I and Chapter II at the 2012 Pacific Meeting of the American Philosophical Association and at the 2012 Bowling Green Workshop in Applied Ethics and Policy. I am grateful for the criticisms and suggestions I received from audience members on these occasions. Michael McKenna, whom I met at Bowling Green, has been tremendously helpful to me and I owe him special thanks. Michael has provided me with good feedback on my work but perhaps more importantly in almost every conversation I have had with him Michael he has been incredibly encouraging. The job market can be quite destructive psychologically and at a time when I was feeling rather pessimistic Michael's kind words and his willingness to help gave me a much-needed boost.

Material contained in the first half of Chapter I and some of the ideas set forth in Chapter II comprise the bulk of a paper titled "Towards a Theory of Interpersonal Manipulation," which will appear in the volume *Manipulation: Theory and Practice*, edited by Michael Weber and Christian Coons and published by Oxford University Press. The second half of Chapter I, in which I discuss the relationship between manipulation and the rational capacities, will appear as "Do Manipulators Always Threaten Rationality?" in *American Philosophical Quarterly*, which is published by University of Illinois Press. I want to thank these publishers for granting me permission to use this material here.

To my parents I of course owe everything. My mother, Luba Gorin, and my late father, Zeev Gorin, always supported my sister and me unequivocally, so long as we did what any reasonable person understands should be done and stopped acting like idiots. Failure in the latter regard led straight to *equivocal* support. I do not believe that either of my parents ever tried to end an argument with their children by telling us, “Because I said so.” Perhaps this is why I seem constitutionally incapable of seeing the appeal in Divine Command Theory. Also, for better or for worse, it helps in contemporary academic philosophy to have a thick skin and I am sure my parents and my sister Shlomit helped me to develop mine. My father probably was the most critical person I have ever known, critical in the sense Marx spoke of his 1843 letter to Arnold Ruge, “referring to *ruthless criticism* of all that exists, ruthless both in the sense of not being afraid of the results it arrives at and in the sense of being just as little afraid of conflict with the powers that be.” I am extremely grateful to be able to report to my mother that I have completed my dissertation. I am profoundly sad that my father is not alive to lovingly berate me for not having finished it more quickly, just in case someone important were to die before it was done.

Nobody has supported me more over the last decade than Acacia Springsteen and nobody has sacrificed as much in supporting me. One of the most gratifying things about completing my dissertation is the sense I have of its having come about as a result of a partnership. Acacia has eyes for things I cannot see. I know I have benefitted enormously as a philosopher, as a partner, as a father, and more generally as a human being as a result of her sharing her perspective on life—and indeed her life itself—with me. Perhaps my greatest fear about not completing the PhD was that I would have squandered years of

Acacia's devotion to my work. I hope with all my heart she feels it was worth it, and I am excited to continue building our lives together such that it will always be true that it was worth it. Thank you, Acacia, for your guidance, your support, and your love. This dissertation is yours as much as it is mine.

Finally, there are the tiny giants, my children Imogen and Lev. I have slept less over the last four and a half years than I did during the preceding two, and although I am awake more of the time I have less of it. It is hard to overstate how much my children helped me write this dissertation, despite weighing a combined total of fifty-three pounds and being totally illiterate. They have given me the best incentive a person could hope for to strive to do well.

Table of Contents

Preface	1
I: Manipulation and Common Wrongs	7
II: Manipulation and the Tracking of Reasons	47
III: The Ethics of Manipulation	96
IV: Manipulation and Libertarian Paternalism	120
<i>Bibliography</i>	154

Preface

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 dissertation lies in distinguishing between the manipulative use of means of interpersonal influence and the non-manipulative use of these means, and in explaining what it is about manipulation that gives us reason to avoid engaging in it.

As a first step toward meeting this difficulty, it will be useful as a preliminary step to classify the various forms of manipulation into a small number of distinct types. Rather than focusing on the means a manipulator might choose to influence the behavior of a manipulee—a project that, I suspect, probably would result in a long list of disparate phenomena—I will distinguish between different sorts of manipulation on the basis of the manipulator's target in the interaction.

In the most general terms, a manipulator aims to influence behavior. I understand 'behavior' quite broadly to encompass overt actions as well as the acquisition of mental states (e.g., a propositional attitude, an emotion of short duration, a mood). With respect

to mental states a manipulator may seek to change a manipulee's beliefs, her desires or other conative states, or her emotions.¹ Thus, mental state manipulation may be *epistemic*, *conative*, or *emotional*. An instance of manipulation is epistemic, conative, or emotional either when the mental state is the end at which the manipulator aims or when the mental state is targeted in order to bring about some overt action. The central feature of mental state manipulation is that it involves a targeted change in the manipulee's beliefs, conative states, or emotions. Of course, some cases of manipulation involve the targeting of more than one mental state.

Sometimes mental state manipulation is not intended to lead to any action on the part of the manipulee. For example, a manipulator may wish to cause anger or sadness in a manipulee because the manipulator is being sadistic, or perhaps as a kind of revenge. Alternatively, a manipulator may engage in epistemic, conative, or emotional manipulation because he believes this change will result in some action. For example, a con artist may accurately describe the abject poverty in which many children live in order to evoke the sympathies of his audience, from whom he is seeking "donations." Action is often the upshot of some combination of belief, desire or other conative state, and emotion, and thus by affecting the latter a manipulator can affect the former. When the manipulator seeks a change in the manipulee's beliefs, I will call the manipulation *epistemic*. When she seeks a change in the manipulee's desires or other conative states, I will call the manipulation *conative*. And when the manipulator targets a manipulee's emotions, I will refer to it as *emotional* manipulation. In some cases, a manipulator may seek an emotional response via a change in the manipulee's beliefs or she may seek an epistemic change via an emotional change. For example, a manipulator might claim to

¹ I will understand a mood as a kind of emotional mental state.

feel depressed with the intention that his interlocutor adopts the belief that he is depressed and, as a result, feels sympathy. This is still a case of emotional manipulation, since it is the emotion—sympathy—at which the manipulator is aiming. Alternatively, a manipulator might play sad music on the stereo, intending this to cause a mood in which the manipulee will be more likely to believe some claim. This is a case of epistemic manipulation, as the goal of the manipulator is a change in the manipulee's beliefs.

Sometimes, though, a manipulator may focus primarily on the environment of the manipulee rather than on her mental states. Given a reasonably accurate picture of an agent's salient mental states it is possible to direct her behavior by structuring her environment so that the interaction of the environment with her present mental states will lead to the desired behavior. For example, research has shown that people exhibit a “status quo” or “default” bias.² When presented with two or more options, one of which is already “pre-chosen” for them (the default) while the others require the agent actively to make an alternative selection, agents tend to favor the default option, even when the non-default options are superior. Suppose an employer who is aware of the default bias prefers that her employees select retirement savings option A rather than option B because, though A is superior from the point of view of the employees option A is more expensive for the employer.³ By making option B the default option so that employees have to “opt out” in order to choose option A the employer structures her employees' environment in such a way as to lead to the behavior she seeks. Her behavior with respect

² Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler, "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias," *Journal of Economic Perspectives*, 5(1), 1991, pp. 193-206.

³ Richard Thaler and Cass Sunstein discuss how employers might paternalistically structure default options to increase savings rates among their employees in *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, 2008. See also their “Libertarian Paternalism is Not an Oxymoron”, *The University of Chicago Law Review*, Vol. 70, No. 4, pp. 1159-1202, 2003

to the employees' mental states is neutral. Their beliefs, conative states, and emotions are left untouched. I will refer to cases like this, where the manipulation does not target mental states but rather the manipulee's environment, as *environmental manipulation*.

The distinction I have drawn between the various kinds of mental state manipulation and environmental manipulation does not take us very far in distinguishing manipulation from other forms of interpersonal influence. This is because non-manipulative influences also target mental states or the environment. However, the distinction is helpful because it abstracts away from the particularities of any particular case of manipulation. Though details do matter, a general account of manipulation must emphasize in more abstract terms the similarities that exist between particular instances that diverge widely in their details. The distinction also allows us to focus more closely on the causal routes between a manipulator's action and the manipulee's resultant behavior.

Chapter Outlines

Chapter I: Manipulation and Common Wrongs

Interpersonal manipulation very often involves harm, deception, autonomy violations, the subversion of the manipulee's rational capacities, or some combination of these common wrong-making phenomena. Thus, it is tempting to analyze manipulation by reference to these things. I motivate accounts of manipulation that make such wrongs essential to manipulation and then argue that such accounts fail.

Chapter II: *Manipulation and the Tracking of Reasons*

Manipulation is best understood as a process of interpersonal influence that deliberately fails to track reasons. The failure to track reasons comes in several forms, depending on the means of influence chosen by the manipulator, her motivations in seeking to influence the manipulee, and the manipulee's motivations in behaving as the manipulator intends that she behave. This account of manipulation captures cases of manipulation alternative views leave out while postulating a unifying property of manipulation that can help explain the attraction of competing views.

Chapter III: *The Ethics of Manipulation*

Manipulators intend either to bring about behavior they do not believe to be supported by reasons or they use means of influence that do not reliably track these reasons when they do aim at reason-supported behavior. Consequently, manipulators deliberately leave their manipulees detached from an important aspect of reality, namely, the considerations that ought to govern their behavior—their reasons. Just as morally right actions that are not motivated by the right reasons are lacking in moral worth, behavior that is not guided by good reasons more generally is lacking in normative worth. By leaving their targets detached from the considerations that ought to govern their behavior, manipulators behave in a way that is *pro tanto* morally impermissible.

Chapter IV: *Manipulation and Libertarian Paternalism*

There is growing interest among governments, policy makers, and health care providers in how they might utilize social scientific research on human decision making in an effort

to help people make decisions that improve their own lives and benefit society as a whole. Research suggests that human decision making can be influenced in predictable ways by normatively irrelevant features of the choice situation, e.g., by the order in which choices are presented. Policies involving such interventions may be manipulative. To the extent that they are, the account of manipulation articulated in earlier chapters can provide a framework for evaluating their ethical permissibility. I argue that libertarian paternalism is morally problematic because it relies on means of influence that are controlling and I suggest that the account of manipulation defended in earlier chapters can help us distinguish controlling influence from noncontrolling influence.

Chapter I

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 of 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.

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.”⁴ 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 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.

⁴ Buss, Sarah, “Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints,” *Ethics* 115 (January 2005) p. 226

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, 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.⁵

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

⁵ 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

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 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.⁶

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.⁷

Here Scanlon voices a concern similar to that expressed by Buss when she says that manipulation can “prevent [a manipulee] from governing herself with an *accurate understanding* of her situation.”⁸ 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

⁶ Scanlon, Thomas, *What We Owe to Each Other*, The Belknap Press of Harvard University Press, 1998, p. 298

⁷ 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 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.

⁸ Buss, Sarah, “Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints,” *Ethics* 115 (January 2005) p. 226

principles are generalizations of Principle M.⁹ 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. Or, a 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.¹⁰

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

⁹ Ibid., pp. 299-322

¹⁰ Rovane, Carol, *The Bounds of Agency*, Princeton University Press, 1998, p78

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

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 engaging in it. 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. It also 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, 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 motivate her to behave in a some 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 some possible action of hers will be detrimental to her interests she probably will be strongly disposed to avoid doing that action. Moreover, 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. Manipulation can be used to *prevent* them from doing so. The “libertarian paternalist” policies proposed by Thaler and Sunstein 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.¹¹ 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 to act in ways that benefit them.¹² 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

¹¹ Thaler and Sunstein, *Nudge* and “Libertarian Paternalism is Not an Oxymoron”

¹² Ibid, p. 1184. I discuss Thaler and Sunstein’s work on nudges in more detail in the final chapter.

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. His action was manipulative but beneficent. 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.

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 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.

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.¹³ An autonomy-undermining theory of manipulation that construes autonomy in this way must insist that manipulators intervene between their manipulees' first-order attitudes and their higher-order attitudes. 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.

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

¹³ See, for example, Frankfurt, Harry, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy*, Vol. 68, No. 1, January, 1971

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.

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.¹⁴ 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

¹⁴ I understand 'approve' rather weakly as a kind of (actual or perhaps even hypothetical) pro-attitude.

salient propositional 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 courses. 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).¹⁵ However, my claim that manipulation is consistent with autonomy does not require that externalist theories are true but only that, if they are true, it is not clear that they can easily rule out manipulation as an autonomy-respecting form of influence.¹⁶

¹⁵ Here I have in mind certain counterexamples to internalism. Mele provides some powerful ones in *Autonomous Agents: From Self Control to Autonomy*, Oxford University Press, 1995

¹⁶ I thank George Sher for pointing out that my arguments regarding manipulation and autonomy can remain neutral on the question of which sort of theory of autonomy—internalist or externalist—is the one we ought ultimately to adopt.

I will take two routes toward 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 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.¹⁷ 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.

¹⁷ Thaler and Sunstein, "Libertarian Paternalism Is Not an Oxymoron", p. 1164. They discuss the same example in *Nudge*, pp. 1-4

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 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 autonomy¹⁸ 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. If this is right, a defender of an externalist theory of autonomy is pushed toward a mixed theory, a theory that incorporates some attitudinal

¹⁸ 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).

condition such as the condition requiring that an agent approve of the processes that lead up 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. It still remains to be argued that mixed theories render manipulation consistent with autonomy. Here I will draw from the 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 the same problem. It will be easier to see this with an example. Take John 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.¹⁹

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

¹⁹ Christman, John, "Autonomy and Personal History," *Canadian Journal of Philosophy*, Vol. 21, No. 1, 1991, p. 11

question arises as to whether the agent resisted the development of these attitudes.²⁰ 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 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.

Manipulation and the Rational Capacities

If there is a dominant view of interpersonal manipulation in the philosophical literature, it is the view that interpersonal manipulation occurs when an influencer intentionally

²⁰ 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.

bypasses or subverts the rational capacities of the person he seeks to influence.²¹ I believe this claim about the nature of manipulative influence draws plausibility from two main sources. First, there is a broad range of cases in which it is true that the rational capacities of the manipulated person are bypassed or subverted. Several such cases will be discussed below. Second, because manipulateness is viewed as a negative character trait the concept ‘manipulation’ is typically understood in a highly moralized manner. Consequently, it may be assumed that forms of interpersonal influence that are generally taken to be morally benign or even exemplary—for example rational persuasion—cannot be used manipulatively. The thought is something like this: if manipulation is impermissible, *pro tanto* or otherwise, while rational persuasion is permissible, then rational persuasion cannot involve manipulation and manipulation cannot involve rational persuasion. Since rational persuasion, which is morally benign or even exemplary, always involves, or just is, engagement with the rational capacities of the agent being influenced, bypassing or subverting these capacities is morally problematic. Therefore, manipulation

²¹ This view has been advanced in one form or another by a number of philosophers. See Baron, Marcia, ‘Manipulateness’, *Proceedings and Addresses of the American Philosophical Association*, Vol. 77, No. 2, (Nov. 2003), p. 50, Beauchamp, Tom and J. Childress, *Principle of Biomedical Ethics*, 6th edn (Oxford, 2008), pp. 133-134, Blumenthal-Barby, Jennifer and Hadley Burroughs, “Seeking Better Health Care Outcomes: The Ethics of Using the Nudge,” *The American Journal of Bioethics*, 12:2, p.5, Cave, Eric, ‘What’s Wrong with Motive Manipulation?’ *Ethical Theory and Moral Practice*, Vol. 10, No. 2, (2007), p. 138, Greenspan, Patricia, ‘The Problem with Manipulation’, *American Philosophical Quarterly*, Vol. 40, No. 2 (Apr., 2003) p. 164, Mills, Claudia, ‘Politics and Manipulation’, *Social Theory and Practice*, Vol 21, No. 1 (Spring 1995) p. 100, Stern, Lawrence, ‘Freedom, Blame, and Moral Community’, *The Journal of Philosophy*, Vol. 71, No. 3 (Feb. 14 1974), p. 74. Thomas Scanlon argues that manipulation is morally objectionable 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...” Insofar as the rational capacities play a central role in helping an agent direct her energies toward aims she has chosen, manipulation subverts these capacities. Scanlon, Thomas, *What We Owe to Each Other* (Harvard University Press, 1998), p. 298

always involves, or just is, the bypassing or subversion of an agent's rational capacities, and this is what renders it morally wrong.

I do not mean to suggest that anyone who believes that manipulation necessarily involves the bypassing or subversion of the rational capacities reasons in the way sketched above. However, I do think that if one were to defend what I will call the Bypass or Subvert (BSV) of manipulation, one probably would want to appeal to cases and to emphasize the differences between typical cases of manipulation and typical cases of rational persuasion. In any case, I do not wish to defend BSV, for I think it is false. In what follows I will argue for this claim. After providing several interpretations of what it is to bypass or subvert a person's rational capacities, I show that each interpretation is consistent with the presence of manipulation in the history of the process that led to the behavior in question. In arguing against BSV, I draw a rather surprising conclusion, which is that one agent may be manipulating another even when the only form of influence she uses is the provision of good reasons or sound arguments.

Before moving on to criticize BSV it is necessary to get clear about its central claim. I understand the claim as a disjunction:

BSV: interpersonal manipulation is a process of influence that necessarily either bypasses or subverts the rational capacities of the person whose behavior is being influenced.

I will begin by addressing the first disjunct—that is, the claim that manipulation *bypasses* a manipulee's rational capacities—and then move on to the second—that is, the claim that manipulation *subverts* a manipulee's rational capacities. In order to assess BSV it will be helpful first 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 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.

Rational Capacities: those capacities that enable agents to assess and revise their beliefs in accordance with the basic canons of logic; to evaluate their epistemic and practical options against criteria generated by their beliefs, values and preference sets; to make adjustments to these beliefs, values, and preference sets in light of new information; and to act in accordance with their judgments about what they have most reason to do.²²

To say that one person bypassed the rational capacities of another may be taken to mean that, in influencing someone, the influencer made use of a means of influence that did not engage the influenced person's rational capacities at all. The sorts of examples that motivate this view usually involve a manipulator who has direct access to the causal mechanisms underlying the behavior of the manipulee. Take, for example, Harry Frankfurt's famous would-be manipulator, Black, who through the use of high-tech gadgetry has the power to control the neuro-physiological goings-on in the brain of Jones such that Black can immediately determine how Jones chooses to and indeed does act.²³ Or consider Alfred Mele's case of Beth, an academic working in a department overseen by a dean who wishes Beth were more industrious. The dean hires a team of very capable

²² This conception of the rational capacities is similar to Eric 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.

²³ Frankfurt, Harry, 'Alternate Possibilities and Moral Responsibility', *The Journal of Philosophy*, Vol. 66, No. 23 (Dec. 4, 1969), pp. 835-837

psychologists who learn what makes Beth tick and then via sophisticated brainwashing techniques directly instill in Beth the mental states that make her a highly motivated and productive scholar while eradicating whatever values or preferences were earlier holding her back from diligently pursuing her work.²⁴

Cases like Frankfurt's and Mele's describe manipulators who make use of means of influence that entirely bypass the rational capacities of the agents whose behavior they wish to control. The rational capacities of the influenced person play no role in the processes that determine her behavior. However, there are many cases of manipulation in which the rational capacities of the manipulee do play a mediating role in the process of influence. Here is one such case:

Election: Jones is campaigning to become President of the United States. He knows he needs substantial support among religious conservatives if he is to have any chance of winning the election. In order to appear more attractive to members of this demographic, Jones regularly invokes Scripture while advocating in favor of his political platform at public appearances. Jones is very skeptical about the existence of God, the truth of Scripture, and all other claims the belief in which constitute (or partly constitute) the religious orientation of the voters to whom he is trying to appeal.

If Jones's use of religious rhetoric plays a substantial role in the explanation of why some religious conservatives vote for him, then it is plausible that he has manipulated these voters and that their voting behavior is a product of his manipulation. Nevertheless, Jones has not bypassed the rational capacities of these voters. In fact, the success of Jones's strategy crucially depends on these capacities. After all, Jones intends his audience to perceive his Biblical references as providing them with a reason to vote for him. He

²⁴ Mele, Alfred, *Autonomous Agents* (Oxford University Press, 1995), p. 145

rightly assumes that given the preferences, beliefs, and values of his audience, their belief that he is in the relevant sense “like them” will motivate them to vote for him. If he did not believe this, if he believed instead that religious conservatives were incapable of recognizing that he was providing them with an apparent reason to vote for him, he would not have appealed to their capacity to make the inferences he intended they make. Thus, *Election* shows that manipulation need not involve the bypassing of the rational capacities.

I believe the natural response to this case is simply to concede that the requirement that manipulation entirely bypass the rational capacities is too strict, and to then focus on whatever features of the case seem troubling. For example, despite Jones’s engagement with voters’ rational capacities it remains true that he intended them to behave in ways they would be unlikely to behave if they were better informed about the features of the options about which they were deliberating. The thought is that although he did engage the voters’ rational capacities, Jones did so in a way that subverted these capacities. Thus, though manipulation need not involve the wholesale circumvention of the rational capacities, it is still open that it involves their subversion.

Subversion as Active Interference

According to one possible conception of what it is to subvert the rational capacities, subversion is best understood as *active interference* with those capacities.

Active Interference Subversion: to interfere directly with a person in such a way as to generate psychological states the presence of which is incompatible with the proper functioning of the person’s rational capacities.

Often, in trying to shape others' behavior manipulators elicit psychological states that are incompatible with the manipulees' ability accurately to represent and assess their situations or to behave in a manner that is consistent with their assessments. The following two cases serve as paradigmatic examples of this.

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.²⁵

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.

In *Theater* the manipulator induces psychological states that impede the manipulee from acting in light of her considered judgments about what she has most reason to do. The manipulee initially decided to go to the play and we may suppose this decision was the product of rational deliberation. After beginning to implement the plan that would allow her to carry out her intention, she finds herself strongly drawn toward another course of action that is inconsistent with her earlier plan, and the new course of action is not one on which she had rationally settled. Her rational capacities—in this case her capacity to act consistently with her judgment about what she has most reason to do—have been subverted by the manipulator's stimulation of her compulsion.

²⁵ Cave, ‘What’s Wrong with Motive Manipulation?’ p. 132

In *Legislation* the psychological states induced by the officials interfere with the ability of citizens to evaluate the rationale for the bill and to assess the full ramifications of its passing. Fearful citizens are likely to assign disproportionate weight to the value of policies they perceive as promoting their safety, and thus by identifying the proposed legislation with the promise of security while simultaneously scaring citizens the officials pervert the deliberations of the citizens whose support (or, more accurately, absence of opposition) they seek.

Commonplace examples like *Theater* and *Legislation* reinforce the view that manipulation is a matter of actively interfering with the rational capacities of the manipulee. There are countless examples of manipulation like this where manipulators “push the buttons” of manipulees, giving rise to psychological states whose effect is to overwhelm the manipulee’s ability to assess and revise her beliefs in accordance with the basic canons of logic, to evaluate her epistemic and practical options against criteria generated by her beliefs, values and preference sets, to make adjustments to these beliefs, values, and preference sets in light of new information, or to act in light of her considered judgments about what she has most reason to do.

However, though manipulators often actively interfere with the rational capacities of the people they are trying to influence, they do not always do so. Sometimes a manipulator will take a more hands-off approach and merely exploit an inherent flaw in the rational capacities of the manipulee. Drawing on recent research in behavioral economics, Cass Sunstein and Richard Thaler describe how an influencer’s knowledge of others’ cognitive biases can help the influencer shape the behavior of those she seeks to

influence.²⁶ To take just one example, research shows that the decisions of medical patients regarding potential treatments can be strongly influenced by the way the information about the outcomes of the treatments are framed.²⁷ The following imaginary (though realistic) example illustrates how this works.

Futile Treatment: Dr. Rasmussin's patient, Mrs. Jackson, is very ill. Mrs. Jackson is ninety years old and, in the judgment of Dr. Rasmussin, has a life expectancy of no more than six months. One of her non-life-threatening ailments is curable but the treatment is very expensive and it requires the devotion of scarce medical resources. Dr. Rasmussin believes that providing this treatment to Mrs. Jackson would be futile as she very likely will not live long enough to enjoy its benefits. Moreover, if Mrs. Jackson gets the treatment, then some other younger or healthier patient who would enjoy its benefits will not receive it. In the judgment of Dr. Rasmussin Mrs. Jackson should not receive the treatment. However, Dr. Rasmussin knows that Mrs. Jackson believes that when it comes to attempts at improving her health and extending her life nothing should be regarded as futile. She is adamant that Dr. Rasmussin should provide the treatment. Dr. Rasmussin has no intention of providing the treatment but does not want unnecessarily to alienate or hurt his patient by expressing his unvarnished opinion about the futility of treating her. Instead, in discussing the matter with Mrs. Jackson the doctor makes use of a particular cognitive bias, sometimes referred to as "the framing effect." Rather than truthfully telling Mrs. Jackson that 90% of the patients who receive the treatment survive, he truthfully tells her that 10% do not survive. Upon learning this Mrs. Jackson judges that the treatment is too risky and decides to "refuse" the treatment.

²⁶ Sunstein and Thaler describe some of these methods in their article, 'Libertarian Paternalism is Not an Oxymoron', *The University of Chicago Law Review*, Vol. 70, No. 4, (Autumn, 2003), p. 1159-2012 and also in their book, *Nudge* (Yale University Press, 2008).

²⁷ Sunstein, Cass and Richard Thaler, 'Libertarian Paternalism is Not an Oxymoron', p. 1161. The paper Sunstein and Thaler cite to support their claim about the efficacy of framing medical outcomes in terms of survival vs. in terms of mortality is Donald A. Redelmeier, Paul Rozin, and Daniel Kahneman, 'Understanding Patients' Decisions: Cognitive and Emotional Perspectives', 270 *JAMA* 72, 73, (1993).

Assuming that in this case the framing effect played a decisive role in shaping Mrs. Jackson's decision—that is, assuming that she would have continued to demand the treatment had Dr. Rasmussin framed the information in terms of survival rates rather than in terms of mortality rates—it is plausible that Mrs. Jackson's decision to refuse the treatment and her remaining positively disposed towards her doctor are products of Dr. Rasmussin's manipulating her.²⁸ That is to say, intuitively Dr. Rasmussin manipulated Mrs. Jackson into “refusing” the treatment and into agreeing to the course of action that Dr. Rasmussin favored.²⁹ Dr. Rasmussin manipulated Mrs. Jackson and yet Dr. Rasmussin did not directly interfere with Mrs. Jackson's rational capacities, at least insofar as he did not stimulate psychological states whose presence is incompatible with or threatening to her ability effectively to deliberate about her options and to act in light of her considered judgments. Thus, if manipulation subverts the manipulee's rational capacities, it must do so in a way that does not require the direct interference with those capacities.

A Narrow Teleological Interpretation of ‘Subversion’

Cases like *Futile Treatment* suggest that if manipulation is to be understood as the subversion of the manipulated person's rational capacities we need a conception of ‘subversion’ that does not entail a manipulator's direct interference with a manipulee's rational capacities. Such a conception would cover cases of active interference but would

²⁸ I leave it open for now whether or not what Dr. Rasmussin did was morally permissible. At this stage in the argument I am concerned with establishing that certain instances of influence are instances of manipulation, and not with establishing anything about manipulation's ethical status.

²⁹ Mrs. Jackson did not really *refuse* the treatment because it was not genuinely open to her to accept the treatment, i.e., it was not being offered to her. Dr. Rasmussin's antecedent decision to refuse to provide the treatment rendered Mrs. Jackson's decision otiose, though of course she did not realize this.

be more inclusive in order to capture other cases where the manipulated person's rational capacities are impeded in some way, but where the presence of the impediment is not something for which the manipulator is responsible.

An account of subversion that focuses on the function of the rational capacities rather than on the etiology of the mechanism that undermines them will capture cases of direct agential interference like those described by Frankfurt and Mele as well as those like *Futile Treatment*, in which the manipulator merely exploits an already-existing cognitive defect. On this view, to influence someone in a way that subverts her rational capacities is:

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

This construal of what it is to subvert the rational capacities explains the judgment that Dr. Rasmussin has indeed influenced Mrs. Jackson via a process that subverted her rational capacities. By providing Mrs. Jackson with information framed in terms of mortality rates rather than in terms of survival rates, Dr. Rasmussin succeeded in getting Mrs. Jackson to make a decision she would not otherwise have made, given her set of beliefs, values, and preferences. Dr. Rasmussin's decided to exploit the framing effect because he knew that Mrs. Jackson's background attitudes would make it rational for her to insist on the treatment. Considering the significant weight Mrs. Jackson places on the value of medical interventions, to decide against such an intervention merely on the basis of how information is presented to her rather than on the substance of that information is plausibly to have behaved irrationally, and thus by targeting one of her cognitive biases

Dr. Rasmussin impeded Mrs. Jackson's rational capacities from fulfilling their function. Given the set of her beliefs, values, and preferences, in medical contexts Mrs. Jackson aims to maximize her chances of improving her health and extending her life. Dr. Rasmussin's intervention undermined her ability to achieve this aim. And because the function of the rational capacities narrowly understood is to help an agent achieve her ends, which ends are products of her set of attitudes, Dr. Rasmussin has subverted Mrs. Jackson's rational capacities.

Though the *Narrow Purpose Subversion* view is an improvement on the *Active Interference* view insofar as the former is able to account for a wider range of cases that intuitively qualify as cases of manipulation, it too succumbs to counterexamples. Some cases of manipulation do not undermine the capacity of the rational capacities to fulfill their function but actually enhance this capacity or even supplant it. For example, a doctor may exploit the framing effect in order to get a severely depressed patient to make the decision that is consistent with the patient's considered judgments, but which is difficult for her to make while in the midst of a bout of depression. In such a case the depression undermines the ability of the rational capacities to fulfill their function while the doctor's focused use of the framing effect *enhances* this ability. Here the framing effect functions as a kind of proxy for the rational capacities.

The *Narrow Purpose Subversion* version of BSV also fails to capture some cases of manipulation in which the rational capacities of the manipulee are in no way inhibited from fulfilling their function. Recall *Lucrative Suicide*. James does not stimulate psychological states that are incompatible with Jacques's ability carefully to reflect upon his attitudes. Nor does James exploit some inherent cognitive bias of Jacques's or

otherwise hinder Jacques's rational capacities from fulfilling their function. Given Jacques's considered beliefs, values, and preferences his action is rendered rational. James does nothing to undermine Jacques's ability to deliberate calmly and clearly about his options or to act in light of his considered judgment about what he has most reason to do.

A Wide Teleological Interpretation of 'Subversion'

Perhaps what the case of Jacques and his conniving cousin shows is not that manipulation need not impede the rational capacities from fulfilling their function, but rather that the function of the rational capacities should be understood in some other way. Thus far I have assumed that the purpose of the rational capacities is 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 a more 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.³⁰ 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 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,

³⁰ Derek Parfit and John Broome, 'Reasons and Motivation', *Proceedings of the Aristotelian Society, Supplementary Volumes* Vol. 71, (1997), pp. 99-146

reasons are tools, and a bad reason can work as well as, or better than, a good one.³¹

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 already mistaking the former for the latter. A person who deliberates on the basis of false beliefs or who makes fallacious inferences will often arrive at mistaken conclusions about what she ought to believe or to do. Thus, an influencer who provides defective arguments or reasons or who exploits the presence of false beliefs or the tendency to reason in a defective manner can fairly be said to subvert the rational capacities of the person she influences. A teleological interpretation of rational capacity subversion that takes a broader view of the purpose of the rational capacities can make sense of the intuition that James has subverted Jacques's rational capacities, and thus it provides a more compelling account of the relation between the rational capacities and manipulation.

But this account of manipulation will not work, either. To see why, notice that 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

³¹ Mills, Claudia, "Politics and Manipulation", *Social Theory and Practice*, v21, Spring 1995, pp. 100-101

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 will use good reasons—that is, reasons that really do justify—when these are more effective at bringing about the behavior at which she is aiming. When a manipulator makes use of good reasons or arguments the justificatory quality of the causally effective reason or argument will be merely incidental for her. Yet it does not follow from this that she knowingly disguises a bad reason as a good one.

A similar objection can be brought against Robert Noggle's account of manipulative action, on which account "manipulative action is the attempt to get someone's beliefs, desires, or emotions to violate [relevant] norms, to fall short of these [relevant] ideals."³² According to Noggle manipulators aim to bring about behavior that falls short of the manipulator's epistemic, conative, and emotional ideals. Insofar as an attempt to get someone to behave in ways that fall short of one's ideals is equivalent to an attempt to get someone to behave in ways one believes to be unsupported by reasons, Noggle seems to be committed to something like the *Wide Purpose Interference* interpretation of BSV.³³

But if manipulators sometimes traffic in what they take to be good reasons, then even the wide teleological interpretation of what it is to subvert the rational capacities

³² Noggle, Robert, 'Manipulative Actions: A Conceptual and Moral Analysis', *American Philosophical Quarterly*, Vol. 33, No. 1 (Jan. 1996), p. 44. The relevant norms are epistemic, conative, or emotional, depending on whether the manipulated state is a belief, a desire, or an emotion.

³³ Though I cannot adequately address them here, I believe Noggle's account runs into problems because he overlooks cases of paternalistic manipulation—that is, manipulation that is aimed at bringing about behavior that the manipulator believes *is* supported by good reasons or, in Noggle's favored terminology, behavior that does not fall short of the manipulator's ideals. Sometimes the upshot of manipulative action is behavior the manipulator believes is reason-supported and consistent with the behavioral norms of the manipulator. Noggle's account cannot accommodate this fact.

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. Here are two counterexamples to the claim that manipulation necessarily involves the subversion of the rational capacities, where the latter are understood according to the *Wide Purpose Interference* definition.

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 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. As it turns out I must permanently leave the country just before the two months are up and consequently never deliver the lie.

Global Warming: Candidate Green is running for President of the United States. In her view, the most significant problem we face today is global warming and her presidential run is motivated *exclusively* by her desire to implement policies that will significantly decrease the quantities of greenhouse gases that are being released into the atmosphere. Green knows she cannot win the election if she openly makes the reduction of greenhouse gases the only—or even the central—plank of her political platform. Green consults with polling experts to determine which issues most exercise voters and she decides in advance that she will adopt the policies that will help her win office, irrespective of whether these are policies that she personally supports. Green adopts the most popular positions—with the exception of her position on greenhouse gases—and after arguing persuasively in favor of these policies goes on to win the election. Most of the citizens who voted for her oppose her environmental policies, which policies she begins aggressively to implement once elected.

In the *Trust Me* case it is plausible that when I give you a sound argument tomorrow or next week I am manipulating you; when you make a good choice due to my having given you sensible advice your choice is (at least partly) the product of manipulation; when you judge that I do not tell you things I believe to be false your judgment has been manipulated. 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 your coming to trust me was a product of my manipulating you by giving you sage advice and good arguments. And yet while I manipulate you I fully engage your rational capacities. The process of influence that leads to your trusting me did not bypass your rational capacities, it did not actively interfere with these capacities, it did not exploit an inherent flaw in these capacities or otherwise hinder these capacities from guiding you toward reason-supported behavior. You had good reason to trust me during those two months as I made every effort to help you form true beliefs and to behave in accordance with the dictates of practical reason. Moreover, whatever reason you might otherwise have had to withhold your trust—namely, that I intended to lie to you—you do not have here, as the state of affairs that generates this reason would not obtain due to my having to leave the country.

Similarly, in *Global Warming* Green engages fully with the rational capacities of voters and she neither actively interferes with the proper functioning of these capacities nor inhibits them in any way from fulfilling their function. We may go further to stipulate that the popular policies were the objectively correct policies, so that the rational capacities of the voters really did guide them toward behavior that is supported by objective reasons. Thus, citizens who vote for Green on the basis of her platform behave

in a reason-supported manner. Nevertheless, it seems that Green acted manipulatively when she constructed a political platform strictly on the basis of its popularity and only because doing so would allow her to implement her favored, unpopular emissions policies.

Global Warming and *Trust Me* are counterexamples to the *Wide Purpose Interference* account of rational capacity subversion. They show that an agent can do what she has good reason to do, to do it in light of those reasons, and yet be manipulated into doing it. Because manipulators care only about the causal efficacy of reasons and not about their justificatory qualities, they will appeal to good reasons when they judge that these will be effective in bringing about the outcome they seek.

The account of manipulation according to which manipulation essentially involves the bypassing or subversion of the manipulated agent's rational capacities is attractive. BSV postulates a unifying property of manipulation that both organizes our intuitions about a wide range cases and purports to explain how manipulation differs from the proffering of good reasons or arguments. However, I have argued that each of several interpretations of BSV is vulnerable to counterexamples, and therefore that the dominant view of interpersonal manipulation is false. Moreover, some of the counterexamples—*Trust Me* and *Global Warming* in particular—reveal an interesting and perhaps rather surprising truth about manipulation: that the provision of reasons and arguments—good reasons and sound arguments—can be used manipulatively. One implication of this result is that insofar as manipulation is thought to be morally problematic, providing others with good reasons and sound arguments can sometimes be

morally problematic. I will turn to the ethics of manipulation in Chapter 3 and Chapter 4.

Before doing so, however, I want to say more about what I take manipulation to be.

Chapter II

Manipulation and the Tracking of Reasons

Toward the end of the last chapter I claimed that sometimes the attitudes or actions manipulees are led to have or to do are supported by good reasons, but that when this is the case it is in some sense a matter of luck. In this chapter I elaborate on and defend this claim. I believe the plausibility of the claim that motivated the last section of the previous chapter—that is, the claim 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.

Manipulation, Reasons, and Luck

There are at least two ways in which manipulation can fail to track reasons. The first involves manipulation that is not motivated by reasons that, with regard to the manipulee, the manipulator believes to be good reasons. I will refer to behavior the manipulator does not believe 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 interests of the manipulee. 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 of or interference with the rational capacities of the manipulee fall into this category. When an agent believes she has reason to do something and is able and motivated to do it, getting her to fail to do what she has reason to do will 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 behaves in ways that are supported by reasons.

Things are more complicated when the manipulator believes the behavior he is seeking from the manipulee is supported by 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 cases as *paternalistic manipulation*. In such cases 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 it furthers those interests practices paternalistic manipulation. The "libertarian paternalist" nudges advocated by Thaler and Sunstein are examples of

reasonable manipulation because the nudger is aiming at behavior she believes will promote the interests of the nudgee.³⁴

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 *Trust Me* case described above 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 instrumental role in my intentions.

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 this 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 best, the reasonableness of the behavior motivates the action

³⁴ Thaler, Richard and Cass Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, 2008. See also their "Libertarian Paternalism is Not an Oxymoron", *The University of Chicago Law Review*, Vol. 70, No. 4, (Autumn, 2003)

of the manipulator indirectly, as when the reasonableness is instrumentally valuable with respect to the manipulator's ends.

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 tend to 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 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. In the case described above 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 in the last chapter, 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 motivated by the narrow instrumental value of that method. A fundamental feature of interpersonal manipulation, one that is reflected in the non-metaphorical³⁵ 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.³⁶ 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 only about "getting it done."

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 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

³⁵ Joel Rudinow distinguishes between metaphorical and non-metaphorical uses of the term. See Rudinow, Joel, "Manipulation", *Ethics*, Vol. 88, No. 4, 1978, p. 339. Kligman and Culver develop this point nicely. See Kligman, Michael and Culver, Charles, "An Analysis of Interpersonal Manipulation," *The Journal of Medicine and Philosophy*, 17, 1992, pp. 173-197

³⁶ One possibility is that in the interpersonal context manipulation involves the use of agents, of persons, as means. I will discuss this possibility later in my more direct discussion of the ethics of manipulation and in particular of the viability of Kantian condemnations of manipulation. 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.

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 try to adjust those means in an effort to ensure that they do lead to such behavior. The capacity of the means of influence to bring about reason supported behavior is a necessary condition for their being chosen and those means will be discarded if the influencer believes they do not have this ability, even if she still believes they will bring about the same behavior. It is the reason-supportedness of the behavior that, for paternalistic manipulators, provides an independent reason 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 because they believe the manipulee has these reasons (perhaps in addition to 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 (or part of one) to bring that behavior about.

Thus, paternalistic manipulation tracks reasons in a way unreasonable manipulation and non-paternalistic reasonable manipulation do not. Nonetheless, 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 unwillingness or inability of the recipient of the influence to behave in the ways the influencer intends that she behave despite their being reason for her to do so. 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 (i.e., non-paternalistic reasonable manipulation and unreasonable 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.

Manipulators are sometime 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 non-manipulative 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—there are no such reasons beyond the instrumental value of the means.³⁷

The claim, scrutinized in Chapter I, that manipulation necessarily bypasses or subverts the rational capacities is false. As I showed, it is possible to manipulate someone

³⁷ There is more to say about the distinction between paternalistic manipulation and rational persuasion. In Chapter 4 I discuss this distinction in more detail. Part of my aim there is to distinguish between means of influence that track reasons (e.g., giving an argument) from those that do not (e.g., exploiting a cognitive bias).

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, as I will argue below, ethically salient connection between manipulation and the role reasons figure in our behavior. 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 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 propositional attitudes and other mental 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 wants the other person to do, the means of influence (sound argument) reliably "aim at" or "link up with" these reasons and the mental states of the person being influenced 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.

Manipulation's failure to track reasons is the ethically salient feature common to all cases of manipulation. It is my view that 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 this more general property. 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.

Reasonable 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 in the last chapter. 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 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 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 (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 motivated by my recognition of the reasons that support the claims I advance and the advice I offer.

Reason-Tracking Processes

Thus far I have claimed that manipulation is a process of interpersonal influence that fails to track reasons and I have provided a rough schema to categorize the various ways in which manipulation displays a breakdown in the tracking of reasons. I have also contrasted manipulation with non-manipulative forms of influence such as rational persuasion by pointing out that in rational persuasion all the salient elements of the interaction “aim at” or “link up with” the reasons that support the behavior of the influenced agent. In this section I will elaborate on these claims by providing a more general account of what it is for a process to track reasons in the sense relevant for assessing manipulation claims.

I will begin by summarizing the innovative work of John Martin Fischer and Mark Ravizza, whose nuanced theory of moral responsibility is centered on the notion of reasons-responsiveness.³⁸ Next, I will distinguish between a process’s failing to be reasons-responsive in the sense specified by Fischer and Ravizza and its failing to track reasons. Though there are some similarities between reasons-responsiveness and reasons-tracking, the concepts are not co-extensive. Nevertheless, both Fischer and Ravizza’s account of moral responsibility and my account of interpersonal manipulation emphasize the role reasons can play in determining the normative status of an agent’s behavior.

³⁸ Fischer, John Martin and Mark Ravizza, *Responsibility and Control*, Cambridge University Press, 1998

Hence, it will be instructive to use (part of) Fischer and Ravizza's account of moral responsibility to motivate the account of manipulation I wish to defend.

Fischer and Ravizza want to show that the truth of causal determinism is compatible with its being the case that agents are (at least sometimes) morally responsible for their actions. They grant that if causal determinism obtains then it is true that when an agent performs some action she could not have done otherwise in the sense putatively relevant for moral responsibility. Nonetheless, Fischer and Ravizza hold that an agent acting in a deterministic universe may be morally responsible for what she does. This is because they deny that the kind of freedom required for moral responsibility is identical to the ability to do otherwise. On their view to meet the freedom condition for moral responsibility agents must exhibit "guidance control" over their actions. Guidance control, unlike "regulative control," does not require that more than one course of action be physically open to an agent, i.e., that when she does X it is true that she could have done other than X.³⁹ Rather, for an agent to have guidance control over what she does requires that the mechanism out of which her action flows is, first, her own and, second, moderately responsive to reasons. The two conditions—i.e., the reasons-responsiveness of the mechanism and the agent's "ownership" of that mechanism—are independent, and as I will explain below this is important for distinguishing between reasons-responsive processes and those that track reasons. I will begin by summarizing Fischer and Ravizza's account of moderately reasons-responsive mechanisms and then move on to discuss briefly the distinction they draw between a mechanism that is an agent's own and one that is not an agent's own.

³⁹ Ibid., p. 31

It is important to note that by ‘mechanism’ Fischer and Ravizza mean nothing more specific or technical than “the process that leads to the relevant upshot” or even just “the way the action comes about.”⁴⁰ So, for example, when an agent does some action as a result of deliberating about her choices—determining and weighing the various considerations that count for and against doing that action—the relevant mechanism is just the agent's faculty of practical reasoning.⁴¹ And when an agent does some action as the result of an evil scientist’s neurological intervention, the intervention is a part of the mechanism.

A mechanism is moderately reasons-responsive when two conditions are met. The first has to do with an agent’s receptivity to reasons, which itself requires that two conditions be met. First, it must be the case that, holding fixed the mechanism that generated the action in question, the agent would have recognized the presence of a sufficient reason to do otherwise than what she did, given the presence of such a reason. Second, the agent “must exhibit an understandable pattern of reasons-recognition,” which means that the pattern of reasons-recognition is minimally rational and grounded in reality. When these conditions are met, the mechanism is *strongly* receptive to reasons.⁴² For example, suppose an agent who suffers from insomnia but is otherwise healthy and happy is deliberating about whether or not to take some drug that is proven to be effective in treating insomnia. Unfortunately, this drug is not recommended for patients with a certain genetic makeup, which this agent has. For such patients the drug is very likely to cause a quick and painless death. If the agent recognizes that this side-effect provides her

⁴⁰ Ibid., p. 38

⁴¹ Ibid.

⁴² Ibid., pp. 69-73

with sufficient reason not to take the drug, then her mechanism is at least weakly receptive to reasons. But suppose the same agent would not view the fact that the drug causes great pain followed by death as a sufficient reason to not take the drug, even though she does still recognize that death without pain is such a reason. Absent some further (unusual) explanation, the mechanism's pattern of reasons-recognition is not understandable. In order to be *strongly* receptive to reasons, the pattern on which the agent's mechanism recognizes sufficient reasons must be “minimally comprehensible, judged from some perspective that takes into account subjective features of the agent [and is]...also...at least minimally 'grounded in reality.’”⁴³ Thus, according to Fischer and Ravizza a mechanism is strongly receptive to reasons when it allows for the recognition of sufficient reasons to do otherwise when such reasons are present and when the mechanism exhibits an understandable pattern of reasons-recognition.

Next, in order to qualify as moderately reasons-responsive the mechanism must be such that the recognition of reasons will (at least sometimes) translate into action spurred on by the recognition of the reasons that support the action. A mechanism that (at least sometimes) translates the recognition of reasons into an action displays weak “reactivity to reasons.”⁴⁴ Suppose it is in fact the case that the insomnia drug would cause the agent to die painlessly and that he recognizes this to be a sufficient reason to avoid taking the drug. This recognition does not translate into action—the agent sees that he has sufficient reason to avoid taking the drug but, perhaps due to weakness of will or a strong urge to escape wakefulness, the mechanism does not move him to avoid taking the drug. This renders the mechanism *not* strongly reasons-reactive, as the recognition of a

⁴³ Ibid., p. 73

⁴⁴ Ibid., pp. 73-76

sufficient reason does not generate action. However, if it were the case that the drug would cause a painful death, the same mechanism *would* issue in action—the agent would not take the drug and this would be the result of his recognizing this reason. In this case, the agent's mechanism is weakly reactive to reasons because there are *some* sufficient reasons the recognition of which would translate into action.⁴⁵

The mechanism of the insomniac described above is not strongly reasons-receptive if the agent fails to recognize that a painful death provides him with a reason to not take the drug, even if he does recognize that a painless death does provide him with such reason (because this pattern of reasons-recognition is not understandable), and thus the agent would not meet the conditions necessary for having guidance control. Strong reasons-receptiveness is a necessary condition for moderate reasons-responsiveness (and thus for guidance control and moral responsibility). However, so long as the agent's recognition of some sufficient reason would translate into action, the mechanism would be weakly reasons-reactive. A mechanism is moderately reasons-responsive, then, when it displays a pattern of recognition of reasons that is understandable and grounded in reality and when, at least sometimes, it translates this recognition of sufficient reasons into action. According to Fischer and Ravizza's theory, the insomniac would exhibit guidance control over his taking the drug—and hence would meet the freedom condition for moral responsibility—if and only if, 1. he would recognize that there is some sufficient reason for him not to take the drug given the presence of such a reason (e.g.,

⁴⁵ Fischer and Ravizza maintain that “reactivity is all of a piece” (p. 73). That is, they claim that if a mechanism would react to some incentive, then it can react to any incentive. Thus, the fact that an agent would not react to reason X does not render him lacking in guidance control (even when the other conditions for guidance control are met), so long as he *would* react to reason Y. Fischer and Ravizza claim that this view about reactivity is “based on a fundamental intuition” and they go on to provide several examples that support this intuition (pp. 73-75). I do not share their intuition on this matter, though for current purposes the extent to which “reactivity is all of a piece” is not relevant.

that it would cause painless death), 2. the recognition of some sufficient reason for him not to take the drug (e.g., that it would cause a painful death) would issue in his not taking it, and 3. the mechanism on which the agent acts is his own. Guidance control requires strong reasons-receptivity but only weak reasons-reactivity.⁴⁶

It is important to note that reasons-responsiveness is determined by the actual properties of the relevant mechanism of the agent. A mechanism is moderately reasons-responsive when it is disposed to recognize and react to sufficient reasons. The truth of reasons-responsiveness claims ultimately is determined by the actual states or properties of the mechanism at the time of the behavior in question, though these properties or states can be defined by reference to counterfactual scenarios. A mechanism M can be moderately reasons-responsive to sufficient reason R, by way of an understandable pattern of reasons-recognition, even if that mechanism has never had occasion actually to respond to R (because the states of the world that give rise to or comprise this reason never obtain). Thus, on Fischer and Ravizza's account an agent who does action A may exhibit guidance control in doing A even if, given the truth of determinism, the world is such that the states of the world that would count as considerations against doing A do not and could not obtain. The crucial point for my current purposes is that although reasons-responsiveness is determined by the actual properties of an agent's mechanisms, it is not determined by the reasons to which the agent, in acting, actually responded. I will return to this point below.

Fischer and Ravizza do not provide an account of how to individuate mechanisms. On their view, when evaluating the extent to which an agent exhibits guidance control over her behavior it is possible to distinguish between the relevant

⁴⁶ Ibid., p. 75

components of the behavior and the irrelevant components by looking at the particular case. So, for example, if a Frankfurt-style neurophysiological intervener (e.g. Black) uses high-tech gadgetry to generate a brain state the upshot of which is that the manipulated agent chooses to rob a bank, the relevant mechanism is the intervention leading up to and including the brain state.⁴⁷ Such an agent would, according to Fischer and Ravizza's account, lack guidance control if that process would not respond moderately to sufficient reason to do otherwise than rob the bank. Though in such a case the brain state is the product of the intentional tampering of another agent, this is not what determines the agent's lack of guidance control. If the same non-reasons-responsive state came about independently, that is, in the absence of any external interference, it would remain the case that the agent lacked guidance control because the mechanism is not moderately responsive to reasons. The interference of the second agent in Frankfurt's case is incidental to the issue of reasons-responsiveness. According to Fischer and Ravizza it does not matter with respect to the reasons-responsiveness condition of moral responsibility how the mechanism was produced (i.e., intentional intervention or “natural” generation). Thus, reasons-responsiveness is a “current time-slice” property, i.e., questions about the reasons-responsiveness of a mechanism can be answered by looking at the properties of that mechanism at one time. The history of the mechanism—how it came about—makes no difference with respect to its capacity to respond to reasons.⁴⁸

⁴⁷ Frankfurt discusses his case of Black in “Alternate Possibilities and Moral Responsibility”, *The Journal of Philosophy*, Vol. 66, No. 23 (Dec. 4, 1969), pp. 829-839.

⁴⁸ Though for Fischer and Ravizza reasons-responsiveness is not a historical notion, guidance control and thus moral responsibility are historical. This is due to Fischer and Ravizza's second condition of guidance control: the “ownership” condition. In order for an agent to be morally responsible for some action the mechanism that leads up to the action must be the agent's own mechanism in addition to its being

I will return to Fischer and Ravizza's account of reasons-responsiveness below, as I believe that the extent to which an agent is responsive to reasons is an important factor in distinguishing between justified manipulation and unjustified manipulation. For now, I want briefly to explain how reasons-tracking differs from reasons-responsiveness.

Reasons-tracking is similar to reasons-responsiveness insofar as both notions refer to the actual properties of states of the processes at the time of the behavior in question. Thus, reasons-tracking and reasons-responsiveness are both current time-slice notions. However, whether or not a process tracks reasons depends on whether or not the process actually is, at the time of the behavior, responding to the particular reasons that support the behavior. Manipulation claims, unlike guidance control claims, refer strictly to the reasons that did (or did not) play a role in a process of interpersonal influence. In determining whether or not a process of influence is a reason-tracking process one must look to the reasons that are actually playing a role in the behavior. It is not enough to ask

moderately responsive to reasons. This is because in some circumstances the manner in which an agent came to have the behavior-underlying mechanism in question intuitively undermines the agent's responsibility for any action that flows from that mechanism, even if the mechanism is moderately responsive to reasons. For example, if Frankfurt's high-tech intervener, Black, brings it about via his sophisticated gadgetry that his victim V does X, it may still be true that the mechanism on which V acted was moderately responsive to reasons. If, given the presence of some sufficient reason not to do X, V would have been receptive to this reason, and if, given the presence of some sufficient reason not to do X (which may or may not be the same sufficient reason) V would have reacted by avoiding doing X, then the mechanism on which V acted in doing X is moderately responsive to reasons. Yet, according to Fischer and Ravizza, V is not morally responsible for doing X, and this is due to the particular history of the mechanism that issued in V's Xing. Intuitively, the mechanism on which V acts is not his own.

Fischer and Ravizza argue that a mechanism becomes an agent's own in the relevant sense when the history of that mechanism includes the agent's having *taken responsibility* for it.⁴⁸ I cannot discuss the details of Fischer and Ravizza's notion of taking responsibility here, and thus will offer only a rough summary. An agent takes responsibility for a mechanism when she views herself as an agent—that is, as a being whose decisions and actions have consequences in the world—and when she views herself as an appropriate target of the reactive attitudes of the other members of her moral community. Finally, her judgments about her status as an agent and as an appropriate target for the reactive attitudes must be based in an appropriate way on the evidence.⁴⁸ By having taken responsibility for a mechanism an agent makes it her own. According to Fischer and Ravizza, then, an agent exhibits guidance control with respect to some action, and is therefore morally responsible for that action, when she acts on a moderately reasons-responsive mechanism for which she has at some point taken responsibility.

whether the agent *would* have behaved differently given the presence of some sufficient reason (although the answer to this question will establish something about the actual properties of the process). Rather, in trying to establish whether or not some interaction is manipulative one asks if the influencer is, in acting, motivated by reasons she believes do support the behavior she is seeking to illicit from the agent she is influencing, if the means of influence she chooses properly link up with these reasons, and if, in behaving as the influencer intends that she behaves, the influenced agent is motivated by these same reasons.

On my account of manipulation, the relevant process always extends beyond the agent whose behavior is being evaluated as a possible product of manipulation. This is the case, obviously, because manipulation involves (at least) two agents—the manipulator and the manipulee. Interpersonal influence cannot be explicated by reference to only one person. Thus, reasons-tracking as I understand it is a process that begins in the agent doing the influencing and ends in the behavior of the agent who is influenced. In order to determine whether or not a process of influence tracks reasons in the sense relevant for assessing manipulation claims one must examine the role that reasons play in the behavior of each agent as well as the relations between the motivations of the agents.

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 she believes 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 she believes support the behavior of

the influenced agent. Non-Paternalistic Manipulators either believe these reasons do exist or are agnostic about their existence.

Paternalistic Reasonable Manipulation: a process of interpersonal influence in which the influencing agent is motivated by reasons she believes 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.⁴⁹

Each of the three forms of manipulation refers to role that reasons play in the actual process of influence. Whether or not the manipulated agent *would* respond to a sufficient reason to do what she does or to do otherwise than what she does, is irrelevant with respect to the question of whether her behavior is the product of manipulation. Reasons-tracking, unlike reasons-responsiveness, cannot be explicated by reference to counterfactual scenarios.

The Manipulation Principle

I believe the failure of manipulation 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.

⁴⁹ In the final chapter I discuss paternalistic manipulation in more detail. One of the main objectives of that chapter is to distinguish between means of influence that do bear a normative relation to the behavior at which they are aimed.

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 is 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 normative considerations, this principle allows us to avoid the controversial conclusion that manipulation is always wrong (or even always *pro tanto* wrong). However, the principle still 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 no or at best only negligible benefits to the patient. This would be impermissible irrespective of the extent to which the patient is reasons-responsive. This suggests that *Manipulation Principle 2* 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 influencer believes the behavior she seeks to bring about is, from the perspective of 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—by stimulating some compulsion of yours, say—while you are severely depressed 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 entails 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 today—it cannot wait until tomorrow. I know that you would be happy to loan me the money if you were not depressed. Indeed, I know that tomorrow you will be happy that you did loan me the money and will recognize the reasons that supported your doing so. 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 influencer believes the behavior she seeks to bring about is, from the perspective of the person whose behavior will be influenced, supported by good reasons.

The addition of “*pro tanto*” 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.

Manipulation Principle 4 provides a basis for distinguishing between morally permissible manipulation and manipulation that is morally impermissible. As such, the principle assumes without argument that there is something morally wrong with manipulation, at least sometimes. Though this assumption is consistent with our ordinary judgments about manipulative influence, much more needs to be said in its defense, as the principle itself does not provide any explanation for what makes manipulation wrong (when it is wrong). Thus, the more general argument in support of *Manipulation Principle 4* comes later, in Chapter 3, where I provide an explanation for what makes manipulation wrong (when it is wrong).

The Disjunctive Account

Thus far I have argued that manipulation is a process of interpersonal influence that fails to track reasons and I have also suggested that this is what renders manipulation ethically suspect. However, one might accept both that manipulation fails to track reasons in the way I have specified and that this renders it different from paradigmatically innocuous or exemplary forms of interpersonal influence without acceding to the further claim that this is what renders manipulation wrong.⁵⁰ Clearly, morally wrong phenomenon X may differ from morally exemplary phenomenon Y with respect to feature Z without its being the case that Z best explain X's wrongness.

There are several routes one might take in responding to this point. First, one might insist that intentionally failing to track reasons when influencing others is always

⁵⁰ I thank Melinda Fagan for pushing this line of argument.

wrong (or *pro tanto* or *prima facie* wrong) and try to defend this claim. I do not believe such a defense can succeed because the claim is simply too strong and would require revising some strong first-order judgments about particular cases. Some cases of manipulation are not even *pro tanto* or *prima facie* wrong. For example, it is not wrong, *pro tanto* or otherwise, to manipulate an armed, insane person into surrendering his weapon. Though there may always be something unfortunate about situations in which persons are manipulated, it does not follow from this that all such cases involve a moral wrong.

Second, one might argue that though it is not always wrong to manipulate others—because it is not always wrong intentionally to influence them in ways that fail to track reasons—it is wrong to influence them in this way when certain background conditions are met. On this view it would remain true that manipulation is usually wrong so long as these conditions are usually met.⁵¹ But in cases where these conditions are not met—for example, in the case of the insane man with a gun—the manipulation is not wrong. *Manipulation Principle 4* is an expression of this view. I believe this approach, which remains committed to the claim that its failure to track reasons is what renders manipulation *pro tanto* wrong, is the most plausible approach to explaining the ethical significance of manipulation. I will say more about it below. But first I must address a third, related possibility, which is that manipulation's failure to track reasons provides nothing by way of an explanation for manipulation's wrongness.

One might think the higher-order property 'failing to track reasons' is normatively inert. The failure to track reasons is indeed a necessary condition of manipulation but it is not sufficient to make manipulation wrong. When manipulation is wrong its wrongness

⁵¹ I thank George Sher for pressing this point.

derives from some other property or properties, for example from the presence of deception, harm, the undermining of autonomy, or the bypassing or subversion of the rational capacities. On this view, if a case of manipulation involved none of these things there would be nothing wrong with it even though it failed to track reasons. And when manipulation does involve some of these things, it is wrong only when they are wrong. If this is right, then there is nothing even *pro tanto* or *prima facie* wrong about interpersonal manipulation *per se*. To explain why we object to manipulation—why calling someone “manipulative” is a form of negative moral appraisal—we need only to understand why we object to the morally wrong means manipulators often employ, e.g., why calling someone “a deceptive person” is a form of negative moral appraisal. This account of manipulation’s wrongness will be a disjunctive one, with the list of the usual suspects (i.e., deception, harm, the undermining of autonomy, the bypassing of the rational capacities, etc.) providing the sufficient condition in the first stage of the analysis of manipulation’s wrongness. Later stages would require explanations for the wrongness of the phenomena comprising the list.

Though I cannot see any way to argue decisively against a disjunctive view, I think there are a number of considerations that count against it as compared to the alternative unified theory I favor. First, assuming that simplicity is theoretical virtue, a unified account that appeals to one general property is preferable, *ceteris paribus*, to some other account that must appeal to a number of disparate properties.

Second, if the general phenomenon—in this case the failure to track reasons—can itself provide an explanation for what is morally objectionable about the particular phenomena that make up the list on the disjunctive account, then the unified theory will

be more explanatorily powerful than the disjunctive one. Just as it is possible to wonder what, if anything, is wrong about the failure to track reasons, it is possible to wonder what, if anything, is wrong with harm, deception, the undermining of autonomy, and the bypassing or subversion of the rational capacities. If it turns out that each of these things shares some characteristic with the others and that this characteristic itself has some independent intuitive force as a wrong-making feature of manipulation, then the disjunctive account will be subsumed by the simple one. Clearly an account that would explain the wrongness of manipulation while also providing a unified view of the wrongness of deception, harm, the undermining of autonomy, and the bypassing or subversion of the rational capacities will be superior to an account that gestures toward the independent wrongness of each of these things. Even if the wrongness of the disjuncts were to be spelled out this would not say anything about the feature or features they share, and hence it will remain unclear on such an account why these features should be lumped together in the analysis of the concept of manipulation. Given two cases of manipulation, each of which involves a different disjunct (e.g., harm, deception), a defender of the disjunctive account must insist on the presence of two distinct wrongs such that the two cases share no normative characteristics (aside from the morally “thin” property of being wrong⁵²). Thus, it would not be clear why an observer might call each of the actions (or the agent) “manipulative” and intend this to express an evaluative judgment. Such a judgment would be off the mark as the manipulation would not itself be the ethically salient feature. According to the disjunctive view the fitting response would be to criticize the first actor for her use of e.g., deception and the second for her e.g.,

⁵² Bernard Williams discusses the difference between “thick” and “thin” ethical concepts in *Ethics and Limits of Philosophy*, Harvard University Press, 1986

causing harm. If manipulation is itself a morally neutral concept, then calling an action or an agent “manipulative” should serve no better as a tool of moral censure than calling it/him an “action” or an “agent.” Clearly common usage understands “manipulation,” “manipulative,” etc. as morally evaluative terms when they are used in reference to instances of interpersonal influence.

But perhaps there is a way to make common linguistic usage consistent with the disjunctive account of manipulation. The idea here would be that ‘manipulation’ is an umbrella concept that covers the more particular illicit forms of influence that are its elements. It functions as the expression of a negative evaluative judgment not because there is something *independently* wrong with manipulation, but simply because there is always or usually something wrong with it, *viz.*, it always involves at least one of the normatively-salient features that, together with the others, makes up the list that forms the core of the disjunctive account. When some instance of interpersonal influence is said to be manipulative, that just means that it is (perhaps among other things) deceptive, or harmful, or autonomy-threatening, etc. On this view, to call an action or agent “manipulative” is to express the same negative moral judgment that would be expressed in the case of deception, the undermining of autonomy, and so on.

Though this response on behalf of the disjunctive view is consistent with ordinary linguistic usage insofar as such usage assumes that ‘manipulation’ carries a negative valiance, it cannot overcome the original difficulty to which it purports to respond. This difficulty, stated another way, is that if the disjunctive account is true, then the proposition ‘X is manipulative’ expresses the same evaluative judgment as ‘X is deceptive’ when the manipulation in question is deceptive manipulation. And the same

holds for any proposition of this form, where the predicate in the second sentence is one of the normatively salient concepts on the disjunctive account's list. To claim an action or an agent is manipulative is on this view just to claim that it/he are deceptive, or harmful, or disrespectful of others' autonomy, and so on. However, the version of the disjunctive view currently under discussion begins with the claim that there is nothing *pro tanto* wrong with the use of means of interpersonal influence that fail to track reasons. Manipulation is *pro tanto* wrong just in virtue of the *pro tanto* wrongness of some of the features it always involves. But if this is right, then it remains mysterious why 'manipulation' should be thought to be a term of moral criticism at all, given that we possess the purportedly finer grained and more informative concepts that make up the disjunctive account.

I believe the account I favor, according to which manipulation fails to track reasons, can explain in substantive terms what deception, the intentional causing of harm, the undermining of autonomy, and the bypassing or subversion of the rational capacities have in common. If I can vindicate this claim, then my account, unlike the disjunctive account, can explain why the disjunctive account initially appears attractive. The disjunctive account does not offer criteria for which wrong-making phenomena should appear on the list that comprises manipulation's ethical elements. Therefore, on this account it remains unclear why we should not also include other forms of interpersonal influence that are also wrong—for example, coercion. As I will argue below, my account can distinguish between manipulation and coercion in a way that makes sense of the list favored by the disjunctive account (which does not include coercion), but it is not obvious that the disjunctive account can do the same. An account of manipulation that

can distinguish between the items found on the disjunctive account's list and those that are not included will be superior to the disjunctive account, for the latter merely provides a list and no criteria for determining the members of the list.

Manipulation and Coercion

The philosophical literature on coercion is too rich to address here in any kind of comprehensive or systematic way. However, it is possible to extract one fundamental feature of coercion that is relatively uncontroversial and which provides the explanation for how on my account coercion differs from manipulation.

When agent P coerces agent Q to do action A, P does so by credibly claiming that Q's failure to do A will lead to a state of affairs that Q judges to be worse than Q's A-ing. As it stands, P's claim need not be coercive, as predictions might also take this form. What makes P's claim coercive is that P indicates that he, P, will bring about the unwanted state of affairs in the absence of Q's A-ing. It is not just that some undesirable state of affairs will transpire on Q's failure to A, but that P will choose to bring this state of affairs about on the condition that Q not A. Coercion differs from prediction most clearly in that unlike someone who merely makes a prediction, a coercer claims that he will bring about some undesirable state of affairs that it is in his power either to bring about or not to bring about, and that whether or not these consequences are realized is in a crucial sense up to the will of the coercer. The coercer does not merely claim that something bad will happen, he claims that he will make something bad happen in the event that and—more importantly—because the coercee fails to behave in the way the coercer intends that he behave. It is an essential part of the interaction that P credibly

assumes both the ability and the *responsibility* for the consequences that P claims will flow out of Q's failure to A. In other words, coercion involves a threat.

The classic example of coercion involves a mugger P who threatens his mark Q that unless Q gives P his wallet P will cause serious physical harm to Q. Here Q is faced with two possible states of affairs. The first is his parting with his wallet. The second is his being seriously physically harmed (and perhaps losing his wallet in any case). Presumably Q prefers the first state of affairs to the second and so will turn over his wallet to P, and he will do so because (or partly because) he believes doing so will make it less likely that P will bring about the second state of affairs.⁵³ We can analyze this case in the terms I have set out in discussing manipulation, that is, in terms of the relations between the motivations of the influencer, the motivations of the agent being influenced, and the way reasons figure in these motivations. In the standard case, P will not be motivated by reasons he believes support the behavior of Q, as there are no such reasons prior to P's making his threat. Once the threat is made, though, there are such reasons, as the threat generates them. P intends that Q recognize that Q now has good reasons to part with his wallet. If he is rational, Q will surrender his wallet in light of the reasons that support his doing so, i.e., in light of the fact that a failure to do so will result in significant bodily harm. In handing his wallet over Q behaves in a way that is supported by reasons and he does so in light of these reasons.

P intends that Q surrender Q's wallet and he intends that Q do so in light of the reasons that support his doing so. However, in coercing Q, P is not motivated by the reasons that support the behavior but is motivated instead by (let us suppose) his desire

⁵³ Nozick, Robert, "Coercion," in *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, Sidney Morgenbesser, Patrick Suppes, and Morton White (eds.), New York: St. Martin's Press, 1969, pp. 440–472

for money. P uses the reason-supportedness of the behavior as a “causal lever”⁵⁴ to get Q to behave in the way he wants him to behave. On the one hand, this looks like a case of unreasonable manipulation, as prior to the coercion it would be unreasonable (let us suppose) for Q simply to give P his wallet. On the other hand, once the threat is made Q’s behavior in surrendering his wallet is indeed supported by reasons and both P and Q recognize this. Thus, subsequent to the threat’s being made the interaction shares features with cases of reasonable manipulation.

I think the best way to distinguish between coercive cases like this and cases of manipulation is by noting that not only is P motivated exclusively by the causal efficacy of the reasons that support Q’s behavior—something that is consistent with P’s committing reasonable manipulation—or that, prior to the manipulation P is in no way motivated by reasons that support the behavior at which he is aiming—something that is consistent with unreasonable manipulation—but that P’s credibly presenting Q with a choice the consequences of which P controls actually generates the reason(s) that supports Q’s behavior. Q’s reason to A just is that P will cause him more harm as a result of his refusing to A than he would face were he to A. That A-ing is attached to greater harm than not A-ing is a fact controlled by P and for which P is responsible. This is why Q’s surrendering his wallet prior to P’s threat would be unreasonable but his doing so in light of the threat is reasonable. P’s credibly threatening Q transforms unreasonable behavior into reasonable behavior by constituting the reason that supports this behavior.

Manipulation does not generate reasons in this way. When manipulation is unreasonable, the manipulator is not motivated by reasons that support the behavior (as

⁵⁴ In her essay “Politics and Manipulation” Mills uses this term to designate cases where the influencer exploits the causally efficacious dimension of reasons but is not interested in their justificatory dimension. I discuss her view in Chapter 1.

the manipulator does not believe there are any such reasons) and the manipulee is either not motivated by reasons (e.g., she behaves compulsively or under the influence of normatively irrelevant emotions, cognitive biases, etc.) or else she is motivated by what she incorrectly judges to be good reasons. When manipulation is reasonable, the reasons that support the behavior are not themselves products of the manipulation but rather exist independently of it. A manipulator may steer a manipulee toward recognition of reasons or she may provide pseudo-reasons that are causally effective in bringing about reason-supported behavior, but the act of manipulation will not itself constitute the reason that supports the behavior. Manipulation is a process that fails to track reasons while coercion is a process that generates reasons.⁵⁵ It will be helpful to contrast the classic case of coercion discussed above with a case of manipulation. Because coerced always behave for reasons that do support their behavior (typically, the failure to do so will result in significant harm to them) the relevant kind of manipulation here is reasonable manipulation because it too produces behavior that is supported by reasons.

For Ticks or for Kicks: P desperately needs heart medication but is \$100 short of the amount he needs to pay for it. P knows that his neighbor Q is financially well-off and will not miss \$100. P knows that Q would give him the money if Q believed the medicine would be effective—Q harbors no ill-will toward P and is a very generous person who views philanthropy both as good in itself and as a way for him to promote his own interests. However, P also knows that Q is extremely skeptical of the medical profession, believing it to be comprised of hucksters and charlatans. Consequently, P has good reason to believe that Q will not give P money if Q believes this money will go to doctors or pharmaceutical companies. However skeptical he is about the clinical use of pharmaceuticals, Q is not at all opposed to their recreational use. Knowing this, P approaches Q and asks him for \$100 “for some pills.” Given his familiarity with Q’s beliefs and habits, P has

⁵⁵ I say more about coercion and its relation to manipulation in my final chapter.

good reason to believe that Q will interpret this request as relating to recreational drug use. Q does interpret the request in this way and gives the money to P, who then uses it to purchase his heart medication.

P is motivated by reasons he believes really do support the behavior he is seeking from Q. P believes, plausibly, that his desperate need for heart medication provides a reason for Q to give him money. Q has plenty of money and his reluctance to give P money for heart medication rests on dubious beliefs about the nature of the medical profession—again, absent these beliefs he would happily give P the money and view his doing so both as a good in itself and as a promotion of his own interests. Thus, *For Ticks or for Kicks* is case of reasonable manipulation.

P makes use of vague terms (i.e., “some pills”) with the intention that Q will incorrectly interpret the meaning of the request. P intends his request to lead to Q’s holding the false belief that P will use the money Q gives him to buy drugs for recreational use and that this false belief will provide (perhaps part of) the motivating reason that will explain Q’s giving P the money. Here the breakdown in the tracking of reasons occurs when Q fails to act in light of the reasons that motivate P’s behavior—that is, the reasons that P believes really do support Q’s behavior. Q gives P money for reasons that in no way motivate P’s interaction with Q, as P knows there are no such reasons.

The central point here is that in *For Ticks or for Kicks*, P’s mode of interacting with Q does not generate the reasons that support Q’s behavior. These reasons exist prior to and independent of P’s attempting to influence Q and are what motivates P to try to

influence Q's behavior. When P coerces Q, however, the reasons that both motivate and support Q's behavior are comprised by the mode of interaction itself.

I have argued that coercion and manipulation differ with respect to the way reasons figure in each mode of interaction. Drawing the distinction in this way can explain why coercion does not and should not appear on the list favored by a disjunctive account of manipulation. The disjunctive account cannot explain why coercion should not be included as a wrong-making feature of manipulation, as on that view coercion's wrongfulness should make it a contender for inclusion along with the other wrongful means of influence that are included on the list. I conclude that these considerations count in favor of my account of manipulation's wrongfulness and against the disjunctive account. Next I show how the common wrongful phenomena discussed in Chapter 1—those included on the list comprising the disjunctive account—can be understood in terms of the failure to track reasons.

Reason-tracking and Common Interactional Wrongs

In Chapter 1 I discussed a number of accounts that sought to explain manipulation's ethical dimension and I argued that none of the common wrong-making features to which these accounts appealed—i.e., harm, deception, the undermining of autonomy, or the bypassing or subversion of the rational capacities—provide a necessary condition for manipulation. I then went on in to argue in this chapter that an account that makes the disjunction of these features central to manipulation is more plausible, but that we have good reason to reject this account, too. The disjunctive account's inability to distinguish between manipulation and coercion provides one reason to reject this account, as do

considerations grounded in the ordinary use of the term ‘manipulation.’ I also suggested that another consideration that would count against the disjunctive view and in favor of the view I am proposing would be if the latter could say in substantive terms what each of the wrongful phenomena has in common with the others. The disjunctive account appeals to the wrongness of each of the disjuncts but does not say what they have in common and thus the most it can say about the feature they all share is that each of them has the property ‘wrongful.’ If my account can provide an explanation for why manipulation is wrong (or *pro tanto* wrong) using a principle that also explains what of ethical significant harm, deception, the undermining of autonomy, etc. have in common, the account gains plausibility as a theory of manipulation and also perhaps tells us something philosophically interesting about other wrongful phenomena.

In Chapter 1 I presented a number of cases of manipulation that served to motivate accounts that I then rejected due to their succumbing to counterexamples. I will now return to some of these cases to discuss them in the context of the account of manipulation that has since emerged. Each case, whether it is deceptive, harmful, autonomy-undermining, or threatening to the rational capacities, displays the failure to track reasons that I have suggested is the ethically salient feature of interpersonal manipulation.

Reason-Tracking and Deception

Manipulators often rely on deception to get others to behave in ways they would not behave were they to hold true beliefs about their situation. In Chapter 1 I presented several cases of deceptive manipulation. The first was *Not Credible*. In this case Henry

undermines the credibility of his colleague Elizabeth by giving her bad information about issues on which he is an authority. When Elizabeth repeats Henry's claims in the presence of other experts in Henry's field, they judge that she is incompetent and unreliable.

My account construes *Not Credible* as a case of unreasonable manipulation. Henry intends that Elizabeth behave in ways that are unsupported by reasons. Knowing that Henry is an expert, Elizabeth assumes that what he tells her about his field of expertise is true or, minimally, consistent with expert opinion in his field. Elizabeth does not wish to appear unreliable or incompetent and, presumably, the experts in Henry's field do not want to make inaccurate judgments about others' competence or reliability.⁵⁶ Elizabeth's making false or incompetent claims is not from her point of view supported by good reasons. She may have sought Henry's expertise specifically to reduce the probability that she would come to believe or express such claims. Thus, Henry aims at behavior that is not supported by good reasons. He also uses means—in this case deception—that do not reliably track reasons.

In the next case, *Synagogue*, David truthfully tells his friend Jack that he, David, saw Susan entering a synagogue. David knows that Jack, who has expressed romantic interest in Susan, strongly prefers to date Catholics. David is also interested in dating Susan and, hoping to increase his own chances with her, he intends Jack to believe—falsely—that Susan is Jewish. David knows that Susan is Catholic and that she was not visiting the synagogue to participate in any religious activity. This too is a case of

⁵⁶ I am not taking a stand here on the relations between desires and reasons. It may be the case that Elizabeth recognizes that she has no reason to appear incompetent and consequently forms the desire not to appear this way (as Scanlan would argue) or, it may be the case that Susan's desire not to appear incompetent *is* (or partly is) her reason for not appearing in this way.

unreasonable manipulation, as the reasons that motivate David's interaction with Jack do not support the behavior at which David is seeking from Jack. If we imagine a slightly different case, one in which there are reasons that support Jack's not pursuing Susan (e.g., David knows that Susan and Jack would make each other miserable), the means David chooses—intentionally withholding relevant evidence—do not reliably track reasons. Thus, both in the original unreasonable version of *Synagogue* and in a slightly different reasonable version, the process of influence fails to track reasons.

In the third case, *Flattery*, Carlos compliments his boss Lucinda for her role in their company's recent restructuring in an effort to gain her good will, which good will he believes may benefit him in light of his recent negative performance evaluation. In this case Carlos does not deceive Lucinda about the content of his claims as he does happen to believe that Lucinda performed brilliantly. However, he does seek to deceive her about his intentions in making those claims, as he would have told her the same things even if he did not believe them. If Carlos did not believe his compliments would positively influence Lucinda's attitudes towards him he would not have approached her. Carlos is motivated exclusively by his desire to get back into Lucinda's good graces and by his belief that complimenting her may help satisfy this desire. If we assume the negative performance evaluation whose effect Carlos is attempting to mitigate was accurate, then the behavior Carlos is seeking from Lucinda is unsupported by reasons. If Lucinda were to become more positively inclined towards Carlos it would be due to unreasonable manipulation as there are no reasons supporting this behavior. But if we assume that Carlos's poor evaluation was the result of a mistake or due to circumstances over which Carlos has no control and that Carlos is, in fact, a very good worker, then Lucinda's

becoming positively inclined toward Carlos is supported by reasons. However, in this case too the process that gave rise to Lucinda's behavior fails to track reasons, as Carlos is not motivated by the reasons that would support Lucinda's behavior. Again, Carlos would have "complimented" Lucinda even if he did not believe a compliment was warranted. He is motivated by his desire to win Lucinda's approval whether or not this approval is justified, and hence the mode of influence in this variation of the story is non-paternalistic reasonable manipulation.

As these examples illustrate, using deception can be an effective way to influence behavior, particularly when in the absence of false beliefs the target of the influence would not behave in the desired way. The reluctance to behave in the way a non-deceiving influencer intends that one behave may take one of two forms. First, an agent's situation may be such that she knows no good reason exists for her to behave as her influencer wants her to behave. Here it will be moot for an influencer to rely on rational persuasion. By providing the agent with a distorted picture of her situation a deceiver causes her target mistakenly to view her situation as one that provides certain reasons where, in fact, there are no such reasons. Second, an agent may be incapable of recognizing or becoming motivated by the reasons that do support the behavior. She may, for example, be overcome with emotion or subject to compulsions that swamp her better judgment. Here deception may be used to stimulate the motivation required to behave in reason-supported ways, even though the reasons themselves will not be providing the basis of the motivation. For example, A may be unable rationally to persuade B, who is depressed, to go to a job interview that will likely result in significant benefit to B, but A may be able to motivate B to do so by falsely telling her that a failure to show up for the

interview will result in, say, a financial penalty. Here the reasons that support the behavior play no role in the motivations of B though her behavior is supported by reasons. Any time A deceives B, A makes it the case either that B's behavior is unsupported by reasons or that the motivations that lead to B's behavior (either A's or B's or both) do not track these reasons. I conclude that deception is a process of interpersonal influence that fails to track reasons.

Reason-Tracking and Harm

Manipulators often harm manipulees, as setting back another's interests can sometimes serve to advance one's own. Causing harm in this way is often wrong. Hence, it may be tempting to identify harm as the feature of manipulation that makes it wrong (or *pro tanto* wrong). But as I showed in Chapter 1, manipulation need not be harmful. Thus, manipulation's being *pro tanto* wrong cannot be explained by reference to harm.

Harmful manipulation can be understood as a process that fails to track reasons. In *Not Credible* and in *Synagogue*, the agents who are manipulated may be harmed, depending on which version of these cases we explore. Clearly Elizabeth may be harmed by the perception that she is incompetent, as Jack may be when he decides against pursuing a relationship with Susan. If we assume that the behavior Elizabeth and Jack adopt is harmful to them, then the kind of manipulation to which they were subjected was unreasonable manipulation. In *Off the Wagon*, Adams manipulates recovering alcoholic Wilson into drinking again, which ultimately causes Wilson to be passed over for a promotion at work (Adams gets the job instead). This is a clear case of harmful

manipulation and one in which the manipulee's behavior is not from his point of view supported by reasons.

Harmful manipulation will always be unreasonable manipulation, at least insofar as 'harm' is understood as the setting back of an agent's all-things-considered interests. An agent's all-things-considered interests include the interests of others when the agent adopts these interests as her own. For example, a parent may make great personal sacrifices for the sake of her children and not regard this as being harmful to her, as her interests extend beyond narrowly self-regarding considerations. If we adopt this more expansive notion of what it is for something to be in one's interests and define 'harm' as the setting back of these interests, then harmful manipulation will be unreasonable by definition. As it can be difficult or impossible rationally to persuade an agent to behave in ways that set back her all-things-considered interests, a dedicated influencer may turn to other means of influence such as manipulation. This explains why manipulation is often associated with harmful behavior—the harmfulness of the sought behavior is what rules it out as an upshot of rational persuasion.

If 'harm' is construed to include only narrowly self-regarding interests, then some cases of harmful manipulation will be reasonable, as sometimes there are good reasons for an agent to behave in ways that set back her narrowly self-regarding interests. On this interpretation of 'harm' a parent who takes on significant debt to fund her child's education harms herself, though she behaves reasonably. Similarly, an agent who donates a kidney to her sibling harms herself and yet her behavior is supported by reasons. If the behavior in these cases were the product of manipulation, then the manipulation would be both harmful and reasonable.

None of the preceding claims regarding harm should be taken as seeking to establish an account of the nature of harm. I have assumed without argument that to harm someone is to set back her interests—whether those interests are strictly self-regarding or more inclusive—and I have sought to show that when a manipulator sets back the interests of a manipulee the process of influence fails adequately to track reasons. In most cases harmful manipulation will be unreasonable manipulation as the manipulator intends the manipulee to behave in ways that are unsupported by reasons. In some other cases and given a narrower construal of ‘harm’ a harmed manipulee may have behaved in ways that are supported by reasons. In either case, the method of influence does not track reasons in the way less problematic forms of influence (e.g., rational persuasion) do. This often explains why manipulation was chosen by the influencer, as reason-tracking processes of influence would be ineffective in getting an agent to behave in ways that are harmful to her.

Reason-Tracking and Autonomy

Marcia Baron has argued that “being manipulative is a vice because of its arrogance and presumption, and because the manipulative person is too quick to resort to ruses, to whining, complaining, threatening, and otherwise wearing the other down, and to exploiting emotional needs or a sense of indebtedness.”⁵⁷ Here I would like to focus on the first two vices Baron lists, arrogance and presumption. About these, Baron says

[t]he manipulative person often takes considerable pleasure in getting her way, engineering outcomes, plotting and scheming, and leading another to make a particular choice without the other realizing that she is being manipulated. Manipulativeness also

⁵⁷ Baron, Marcia, “Manipulativeness”, Proceedings and Addresses of the American Philosophical Association, Vol. 77, No. 2, Nov., 2003, p.50

involves arrogance, manifested in at least two ways: in her supposition that others' decisions are for her to make, and in the presumption (in the case of paternalistic manipulateness) that she knows the other's needs, priorities, and weaknesses better than he does.⁵⁸

Though Baron does not discuss the relationship between manipulation and autonomy, anyone who believes that autonomy is a fundamental moral concept probably would find the manipulative behaviors she lists to be ethically problematic. When one agent “engineers” an outcome involving other agents or supposes that she, the influencer, has the right to control others’ choices, or when she “leads another to make a particular choice without the other realizing she is being manipulated” it is plausible to think the behavior of the influenced agent is in some way non-autonomous. Behavior that is “engineered” or which is the upshot of intentionally hidden motives may not be fully self-governed. Thus, Baron’s virtue-theoretic description of manipulation’s wrongness can be recast in terms of manipulation’s autonomy-undermining potential. In Chapter 1 I argued that manipulation is consistent with the preservation of autonomy and that this is true on both internalist and externalist conceptions of autonomy. Here I will show how worries about the autonomy-undermining potential of manipulation can helpfully be understood by reference to breakdowns in the tracking of reasons.

According to the internalist account of autonomy an agent behaves autonomously if but only if her second-order attitudes align properly with her first-order attitudes. If an agent wants to X, does X as a result of wanting to X, and wants to want to X, then her X-ing is autonomous.⁵⁹ The coherence of higher-order attitudes with lower-order attitudes is what distinguishes autonomously-held attitudes from non-autonomous attitudes and

⁵⁸ Ibid., p. 49

⁵⁹ Frankfurt, Harry, “Freedom of the Will and the Concept of a Person”, *The Journal of Philosophy*, vol. 68, No. 1 (January 14, 1971), pp. 5-20

actions are autonomous to the extent that they flow from autonomous attitudes. One of the main objections to this account of autonomy is that an influencing agent can arrange the attitudes of the influenced agent in a way that intuitively renders the attitudes non-autonomous. For example, suppose an agent wants to X but does not identify with this desire, so that the desire to X is not autonomous. A second agent could render the attitude autonomous by manipulating the higher-order attitude so that it coheres with the lower order attitude. Or, she might manipulate the lower-order attitude so that it coheres with the higher-order attitude. Though this would render the attitude autonomously on the internalist account, intuitively the influencing agent is controlling the behavior of the influenced agent in a way that undermines her self-governance. Thus, many cases of manipulation will not count as autonomy-undermining on the internalist account of autonomy. Nonetheless, some will and it is worth discussing how these cases display a breakdown in the tracking of reasons.

If an advocate of the internalist account of autonomy regarded manipulation as wrong due to its purportedly undermining the manipulated agent's autonomy, this would most obviously apply in cases where a manipulator alters the manipulee's first-order attitudes so that they do not cohere with her higher-order attitudes. *Off the Wagon* provides a nice example of such manipulation. Here Adams works to strengthen Wilson's first-order desire to drink alcohol or perhaps to draw Wilson's attention to this desire in a way that will make it more salient and, consequently, more effective in determining his action. As a result, Wilson acts on a desire he does not endorse and, consequently, behaves non-autonomously. By stimulating Wilson's first-order desire to drink alcohol Adams sees to it that Wilson's first-order attitudes diverge from his second-order

attitudes. Considering that Wilson's second-order attitude is supported by reasons, the stimulation of Wilson's first-order desire to drink is unsupported by reasons, as this first-order desire works against the satisfaction of the second-order desire. In manipulating Wilson, Adams seeks behavior that is unsupported by reasons. Where an influencer stimulates an agent's first-order attitudes and where these attitudes are not endorsed by the agent's second-order attitudes the influencer will be judged by internalist theories of autonomy to be undermining the influenced agent's autonomy. Most of the time the behavior that flows from the first-order desire or that desire itself will not be supported by reasons as the failure of the first-order desire to cohere with higher order desires is due to the agent's adopting a higher-order attitude after judging there to be no good reason to endorse the first-order attitude.

According to the mixed externalist account of autonomy defended by Christman the history of an attitude and not only its relation to higher-order attitudes determines whether that attitude is autonomously held. Here again are Christman's autonomy conditions:

- 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.⁶⁰

⁶⁰ Christman, John, "Autonomy and Personal History", *Canadian Journal of Philosophy*, Vol. 21, No. 1 (March, 1991) pp. 1-24

On Christman's view if one or more of these conditions are not met the behavior in question will not be autonomous. By looking closely at what is involved when these conditions are not met we can see they capture something important about the way reasons figure in the behavior of the agent. Christman's conditions require that the agent must upon minimally rational reflection which is free of self-deception (either actual or hypothetical) endorse the development of the attitude in question. The question we must ask here is: How exactly does the failure rationally and reflectively to endorse the history of an attitude threaten the extent to which that attitude is autonomously held? It seems to me the best way to answer this question is by reference to the role that reflection and rationality play in helping an agent recognize what reasons there are. An agent who attends to the developmental process leading to her holding some attitude will resist that process when she judges there to be no good reason for her to hold that attitude or, more generally, for her to adopt attitudes produced by that process. If the agent does not resist the development of the attitude due to a failure to reflect on the way her attitudes are developing or to some failure of rationality (e.g., self-deception), but would resist the attitude were she rationally to reflect on its development, this means that the holding of the attitude or the general acceptance of attitudes produced by that process run contrary to her reflective rational judgment about the reasons she has (or does not have) for holding that attitude or endorsing the process. In other words, she judges or would judge that she has no good reason to hold this particular attitude or, more generally, to hold any attitudes produced by the process of development upon which she is reflecting. If she judged that her holding of the attitude or her endorsing the process that leads to it were supported by reasons, she would not resist the development of the attitude.

It is plausible to view insistence on the presence of rational reflection and on the absence of reflection- or rationality-undermining factors as serving to guard against the adoption of attitudes that the agent would, were she capable of seeing things clearly, judge to be unsupported by reasons. Thus, it seems that Christman's autonomy conditions function to rule out processes that an agent judges not to be supported by reasons. If this is right, then Christman's emphasis on an agent's endorsement of the history of her attitudes can be understood as being motivated by concerns about the extent to which the processes that make up this history track reasons. An agent who does not rationally and reflectively judge her attitudes or the processes that lead to her attitudes to be supported by reasons will not be motivated by such reasons.

Reason-Tracking and the Rational Capacities

I argued in Chapter 1 that manipulation need not bypass or subvert the rational capacities. That being said, manipulation often does bypass or subvert the rational capacities. When it so does this can be understood in terms of manipulation's failure to track reasons. A process that bypasses or subverts an agent's rational capacities will, by definition, fail to track reasons. If, on the one hand, rationality has independent normative force, then an agent always has reason to behave rationally and consequently by bypassing or subverting the rational capacities a manipulator undermines the ability of a manipulee to behave in ways this reason supports. When a manipulator influences an agent to behave irrationally—that is, in ways that are inconsistent with the manipulee's set of preferences, beliefs, and values—this behavior will run contrary to reason. For example, a

manipulator may stimulate a compulsion to produce behavior that is not in keeping with the agent's settled preferences.

On the other hand, if rationality derives its normative force from its relation to whatever reasons there are, irrespective of whether these reasons support the agent's current attitudes, then by bypassing or subverting the rational capacities the manipulation will sometimes undermine the source of value from which the value of rationality is derived—that is, the reasons at which rationality aims. This will be the case when the manipulation in question is either reasonable or unreasonable. Unreasonable manipulation aims at behavior that is unsupported by reasons and thus on the assumption that the purpose of the rational capacities is to link up with reasons, manipulation of this kind subverts the rational capacities. In cases of reasonable manipulation, where the behavior being sought is supported by reasons, the means chosen to influence this behavior do not reliably or properly track reasons, though in the particular case the means do lead to reason-supported behavior. Here the normatively relevant means of influence (i.e., engagement with the reason-directed rational capacities) are bypassed or subverted. For example, a manipulator may exploit the status quo bias to get a manipulee to act in a way that is supported by reasons, though the status quo bias is not one of the elements that make up the rational capacities.

Thus, when manipulation bypasses or subverts the rational capacities, it does so through its failure to track reasons. I conclude that like concerns about the role of deception, harm, and the undermining of autonomy, concerns over the way manipulators fail adequately to respect an agent's rational capacities are concerns about manipulators influencing manipulated agents' behavior via processes that do not track reasons.

Chapter III

The Ethics of Manipulation

In this chapter I elaborate on earlier claims regarding the moral impermissibility of manipulation and then go on to defend an account of what makes manipulation wrong (when it is wrong). I begin with a recapitulation of the analysis of manipulation articulated in Chapter II and the claim I advanced there that it is *pro tanto* morally impermissible to manipulate a reasons-responsive agent. I then provide examples meant to motivate the judgment that manipulation *is* wrong even in cases where the manipulator intends the manipulee to behave in ways that are supported by good reasons. By emphasizing the way the motives of manipulators refer (or fail to refer) to the reasons they believe support the behavior they are trying to induce in the manipulee, I show that manipulators deliberately leave their manipulees detached from an important aspect of reality, namely, the considerations that ought to govern their behavior—that is, their reasons. To support this analysis of the wrongness of manipulation I turn to Julia Markovits’s compelling account of what gives an act *moral worth*. By extending her account of the moral worth of right actions to behavior more generally, I develop a notion of *normative worth* and show how manipulation aims at behavior that is lacking in this kind of value.

In Chapter II I argued that interpersonal manipulation is best understood as a process of influence that deliberately fails to track reasons. I described several different forms this failure might take: sometimes manipulators seek to bring about behavior they believe to be unsupported by reasons; other times, manipulators do believe the behavior

they seek to bring about is supported by reasons, but in these cases either this belief does not motivate them or, when they are motivated by the reason-supportedness of the behavior, they make use of means that do not track reasons. I then advanced the following principle:

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 influencer believes the behavior she seeks to bring about is, from the perspective of the person whose behavior will be influenced, supported by good reasons.

Manipulation Principle 4 provides a basis for distinguishing between manipulation that is *pro tanto* impermissible and manipulation that is not, but it does not explain what it is about manipulation that renders it impermissible (when it is). Up to this point I have asserted without argument that manipulation is sometimes *pro tanto* morally impermissible. While this assertion probably accords with ordinary evaluative judgments much more needs to be said in its defense. For even if ordinary judgments are correct—that is, even if it is true that manipulation is sometimes *pro tanto* impermissible—we are still in need of an explanation of what makes these judgments correct. The central objective of this chapter is to provide an account of the ethical dimension and normative significance of manipulation. An explanation of what makes manipulation *pro tanto* wrong (when it is so) will serve to support Manipulation Principle 4 and consequently to provide a basis for assessing the moral status of manipulation cases.

2. Motivating the Judgment that Manipulation is *Pro Tanto* Wrong

When some act of manipulation is criticized or when a person is criticized for acting manipulatively it is usually the case that the manipulator caused the manipulee to behave in a way the manipulator regarded as being unsupported by good reasons from the point of view of the manipulee.⁶¹ People tend to resort to manipulation when more direct means of influence like rightly-motivated rational persuasion will be unlikely to succeed and such means will be particularly unlikely to succeed when the person to be influenced is in a position to recognize that she has no good reason to do what the manipulee intends that she do. In such cases of unreasonable manipulation the explanation of manipulation's wrongness will be relatively straightforward, as it plausible to think that it is wrongful, or a sign of a character lacking in virtue, deliberately to get others to behave in ways one regards to be unsupported by reasons from their point of view. Consequentialists, deontologists, and virtue theorists may differ about why such acts or the people who perform them are fair targets of moral criticism, but it is fairly easy to see how the various moral theories would address this sort of manipulation.

The ethical significance of reasonable manipulation is more puzzling because here the manipulator intends the manipulee to behave in ways she, the manipulator, regards to be supported by reasons from the perspective of the manipulee. In cases of successful reasonable manipulation the manipulee does what the manipulator believes she has good reason to do and thus any objection to the way she is treated cannot be grounded in concerns over the deliberate hindering of her interests or the manipulator's wholesale failure to take seriously the demands of reason. And yet there does sometimes seem to be something objectionable about reasonable manipulation. This claim can be motivated by

⁶¹ In the terminology introduced in the previous chapter, this kind of manipulation is *unreasonable manipulation*.

comparing cases of reasonable manipulation to cases of non-manipulative reasonable influence like rightly-motivated rational persuasion. In the latter sort of case, the influenced agent behaves in ways the influencer believes to be reason-supported, which reason-supportedness motivated the influencer, and the influenced agent does what she does in light of those reasons. In the former case, the manipulee either behaves in light of considerations the manipulator does not regard to be good reasons or the manipulator is not motivated by these reasons. For example, suppose person P wants person Q to do action X. Person Q is relevantly and decisively responsive to reasons and P believes Q has good reason to do X. P can either convince Q to do X by citing the reasons that support Q's doing X or P can, say, stimulate some compulsion of Q's or engage with the preferences of Q in such a way as to get Q to do X but not for reasons P regards as good. If P can influence Q to do some reason-supported behavior in one of these two ways, it seems preferable (*ceteris paribus*) from the moral point of view for P to choose the non-manipulative process. We need an explanation for why this is so, but if it is true then there is at least a *prima facie* case for the truth of Manipulation Principle 4. Here is a less formal example.

Airport 1: Larry arranges to ride to the airport with William. At the last moment William becomes ill and thus Larry is in need of a ride from someone else. He knows that Katherine very likely will give him a ride if he explains to her both the importance of his getting to the airport on time and the circumstances leading up to his asking her for a ride on such short notice. He also knows that Katherine will very likely give him a ride to the airport if he simply makes his request with no explanation, but only if he asks her immediately after subtly reminding her of the time she inadvertently embarrassed him at a faculty meeting, an occasion about which Larry knows Katherine feels much guilt. Larry has long stopped being bothered by what Katherine did at the faculty meeting and does not believe

that her embarrassing him provides her with a good reason to give him a ride. Larry describes his situation to Katherine, explaining the importance of his getting to the airport, the suddenness of William's illness, and so on, and then asks her for a ride. Katherine agrees to drive Larry to the airport and does so as a result of recognizing the reasons that speak in favor of her doing so, the very reasons that motivated Larry to make his request.

Airport 2: This case is identical to *Airport 1* except that here Larry offers no explanation of why he needs to get to the airport or why he is making his request on such short notice. Instead, he first elicits Katherine's feelings of guilt and then simply asks her for a ride. Katherine agrees to drive Larry to the airport and she does so because she feels guilty about embarrassing him at the faculty meeting.

I think it should be relatively uncontroversial that Larry's behavior in *Airport 2* is *pro tanto* wrong, though it may be all-things-considered morally permissible depending on other details of the case. But absent very unusual facts about Katherine, Larry, or some other features of the situation, Larry's behavior in *Airport 1* gives us no reason—*pro tanto* or otherwise—to criticize him morally. Granting that Katherine's driving Larry to the airport is supported by the same reasons and to the same extent in both cases the moral difference between the cases must be grounded in the process leading up to (and perhaps including) her driving him to the airport. According to the account of manipulation I have been articulating the relevant difference between *Airport 1* and *Airport 2* lies in the way the process of influence tracks, or fails to track, reasons. To explain why Larry's behavior in *Airport 2* is morally problematic while his behavior in *Airport 1* is not we should first examine more closely Katherine's actions in the two cases. If it turns out that there are important differences in what she does or in how she does what she does, then an

explanation of these differences may help illuminate important differences in Larry's behavior, insofar as what Larry does directly shapes what Katherine does.

In *Airport 1* Katherine's motivating reason—i.e., the reason that explains why she did what she did—in driving Larry to the airport coincides with Larry's motivating reason in making his request. Larry asks Katherine to drive him to the airport because her doing so would allow him to catch an important flight and Katherine agrees to drive Larry to the airport for this same reason. The motivating reason—that driving Larry would allow him to catch an important flight—also coincides with what Larry takes to be the reason that *justifies* Katherine's driving him to the airport. That is to say, the features of the situation that Larry believes justify Katherine's driving him—namely, that her doing so will allow him to make an important flight, that he has no other way to get there, etc.—are the same features that both motivate him to make his request and that motivate Katherine to drive him.

It is harder to specify precisely the content of Katherine's motivating reason in *Airport 2*, though it is psychologically plausible that it would involve her feeling a sense of obligation to compensate Larry somehow for having previously caused him embarrassment. Katherine may feel responsible for having embarrassed Larry and if so her motivating reason—the reason that explains her behavior—is that doing so will allow her to “make it up to him” or simply to serve as an indication to Larry that she does not harbor ill will towards him. Larry, however, does not believe that Katherine owes him any favors for having embarrassed him and does not need Katherine's reassurance regarding her attitudes toward him (perhaps he knows her to be a kind and sensitive person who did not intend to harm him). In short, he does not think that Katherine's guilt,

or the causes of this guilt, are considerations that speak in favor of her driving him to the airport on such short notice. Consequently, if the content of Katherine's motivating reason were given by some set of facts about how she made Larry feel at the faculty meeting, this reason would not coincide either with Larry's motivating reason in making his request or with what he takes to be the normative reason that supports his request and Katherine's acceding to it.

In *Airport 1* the process of influence tracks reasons, as Larry explicitly appeals to the reasons he believes support Katherine's driving him to the airport, he is not motivated exclusively by the narrow instrumental value of his citing these reasons as a means of influencing her, and Katherine's recognition of these reasons grounds her motivation to drive Larry to the airport. In *Airport 2* the process of influence does not track reasons, as Katherine's feelings of guilt are not themselves the reasons Larry believes support her driving him, as he chooses these means solely on the basis of their causally efficacious features.

The *Reason to Act for Reasons* Principle

The next question we should ask is whether there is something problematic in general with deliberately getting an agent to behave in ways one believes are indeed supported by reasons, but where the influenced agent does not do what she does *in light of* these reasons but rather in light of what the influencer takes to be some other, weaker reasons or for reasons the influencer regards to be bad. I think the following principle can help make sense of why we judge Larry's behavior in *Airport 1* to be morally superior to his behavior in *Airport 2*.

The Reason to Act for Reasons Principle (RARP): If an agent has sufficient reason to Φ , then that agent has sufficient reason to- Φ -for-that-reason.

Let us suppose that Larry believes Katherine has sufficient reason to drive him to the airport. The reason is that doing so will allow him to catch his flight. *That it will allow Larry to catch his flight* is on Larry's view the consideration that counts in favor of Katherine's driving him to the airport. In *Airport 1*, this is the consideration in light of which Katherine drives Larry to the airport. Thus, Katherine's driving Larry to the airport in *Airport 1* is consistent with RARP—not only does Katherine do what Larry believes she has sufficient reason to do but she does it because she recognizes and responds to this reason. Katherine's motivating reason coincides with what Larry judges to be her normative reason. In *Airport 2*, however, Katherine merely does what Larry believes she has sufficient reason to do. She fails to do what she has sufficient reason to do *because* she has this reason. So here Katherine's motivating reason fails to coincide what with Larry judges to be her normative reason. In *Airport 2* Katherine gets “normatively lucky” when she drives Larry to the airport, as the reason-supportedness of her behavior plays no role in grounding her motivation to drive him.

RARP is a modest principle. What it states is that reasons, understood as considerations that count in favor of behaving in some way,⁶² are behavioral guides. Though the principle is a modest one it is not trivial, for it might be thought that so long as a person does what she has sufficient reason to do—that is, so long as there are considerations that count decisively in favor of her behaving in some way—it does not matter whether or not this reason plays a role in determining her behavior. On this view

⁶² Scanlon, p. 17

so long as one does what one has sufficient reason to do, the demands of reason are satisfied. What matters is that one's behavior accords with reason, irrespective of whether or not the behavior is guided by reason. I will call this view the *Reasons Endorse* (RE) view, as it holds that reasons endorse behavior but do not prescribe it. If RE is true, then RARP cannot help explain the ethical significance of manipulation. Below I will argue that RE is false. But first I want to discuss another related objection to RARP. The objection is that there are cases where an agent has a reason to Φ but lacks any reason to Φ -for-that-reason. Take the following case.

Interview Anxiety: Gabriel is in need of employment. His friend Simon has a good job and speaks to his boss about Gabriel, who has just the sort of experience Simon's boss is looking for in a new employee. Simon convinces his boss to agree to meet with Gabriel for a job interview. Given his dire financial situation Gabriel has good reason to meet with Simon's boss. The problem is that whenever Gabriel reflects on his finances he becomes paralyzed with anxiety. Consequently, if Gabriel were to associate his financial problems with his holding a meeting with Simon's boss he would not make it to the interview.

If his consideration of the reason that supports his going to the interview would undermine Gabriel's going to the interview, then Gabriel might be best off going to the meeting but only in light of some other reason (or for no reason at all). For example, Simon might invite Gabriel to join him at his office for a holiday party and use this opportunity to introduce Gabriel to his boss, so that in preparing to go to Simon's office Gabriel will be relaxed, not anticipating that he soon will be discussing financial matters. Thus, it might seem that in this case Gabriel has a sufficient reason to meet with Simon's boss—namely, that doing so is likely to result in financial stability—but no sufficient

reason to do so in light of this reason.⁶³ If this is right, then *Interview Anxiety* is a counterexample to RARP.

I think this objection gets something right but also that it gets something wrong. The objection is right insofar as Gabriel's anxiety makes it the case that he has good reason to avoid thinking of his meeting with Simon's boss as being related in any way to his financial problems. After all, if he does conceive of the meeting as being financially relevant he will be debilitated by anxiety and likely will miss the meeting, and he has sufficient reason not to miss the meeting. However, the objection is misguided insofar as it suggests that Gabriel's anxiety somehow makes it the case that he now lacks sufficient reason to make the meeting happen and to do so in light of the fact that it likely will improve his financial situation. Gabriel's anxiety makes him practically irrational—his psychological condition renders him unable to do what he has sufficient reason to do—but this does not mean that he no longer has sufficient reason to do it. Gabriel's recognition that his meeting with Simon's boss is likely to lead to financial stability ought to guide his will toward attending the meeting. That it would not do so does not speak against this consideration as behavioral guide, but only against Gabriel's capacity to respond to this consideration in the appropriate way.

Consider that if Gabriel's anxiety were more severe, such that he could not hold a job even if one were offered to him, this would not make it false that he has sufficient reason to take a job. In this case, Gabriel would have good reason to make efforts to reduce his anxiety but he would still have good reason to take a job. In other words, that having a job would undermine Gabriel's ability to carry out the responsibilities required

⁶³ I wish to thank Baruch Brody for raising this objection.

of holding a job would not make it false that he has good reason to carry out those responsibilities and to hold a job. Similarly, Gabriel in *Interview Anxiety* has good reason to take steps toward reducing his anxiety, but he still has good reason to meet Simon's boss and to do so in light of the fact that this meeting will make him better off financially. So, though due to his anxiety Gabriel has reason to avoid thinking of his financial problems when he considers the interview with the boss, it remains true that he has reason to attend the interview and to do so in light of the fact that it will solve his financial problems.

More generally, it cannot be right that reasons endorse behavior but do not prescribe it. Recall that according to RE it does not matter whether or not an agent does what she has sufficient reason to do in light of that reason; what matters is that she does what she has sufficient reason to do. To see why this view is mistaken consider a person who is not responsive to reasons but who nevertheless exhibits outward behavior that is identical to the behavior of someone in her position who is ideally responsive to reasons. Imagine a severely mentally ill person—I will call her Deludia—who suffers from all-encompassing delusions. Her mental states utterly fail to represent reality. However, through sheer happenstance Deludia's mental states correspond (non-representationally) with reality in a way that leads her to behave in reason-supported ways. For instance, Deludia goes to the doctor whenever she has a very high fever but only because this is when she happens to “remember” that medical treatment will cause her hair to turn red, something she welcomes because she believes only red haired people can avoid detection by CIA agents. There is no causal connection between her high fevers and her “remembering” how to avoid detection by the CIA—the etiology of the psychological

processes that lead to this belief and the etiology of the fevers are unrelated. Suppose that Deludia's behavior *always* has this character, so that on the one hand she consistently does what she has good reasons to do but on the other hand she is never motivated by these reasons. An external observer might conclude that she is a robustly reasons-responsive agent while her own explanations of her behavior make reference to all manner of fantastical entities and events. Deludia's behavior runs counter to what RARP prescribes, for though she behaves in reason-supported ways she does not do so in light of these supporting reasons.

Clearly there is something unfortunate about Deludia's situation. Her relations with the world are deeply defective. She is detached from reality, radically mistaken about various features of her environment and their salience with respect to what she has reason to do. Most obviously, she holds many beliefs that are false: There are no CIA agents following her. Medical treatment will not turn her hair red. Having red hair would likely not aid in her efforts to escape detection even if it were true that she were being followed by the CIA. But Deludia's situation is an unusual one insofar as her false beliefs about the world do not cause her to behave any differently than she would behave if she were ideally informed and rational. Due to the way in which her beliefs non-representationally correspond with reality her behavior is always supported by reasons. That is, despite her deeply flawed perspective on the world Deludia always acts in accordance with what she has good reason to do (e.g., she seeks medical treatment when she is ill). Again, from the perspective of an external observer her behavior appears to be consistently, systematically responsive to reasons.

Though her distorted view of the world makes her (or her mental states) defective from an epistemic point of view, if Deludia's behavior would be judged by an external observer to be responsive to reasons and if she does in fact always do what she has good reason to do, it is more difficult to explain what is wrong with Deludia *qua* practical agent. Though many of her beliefs are false she holds true beliefs about which of her possible courses of action are supported by reasons (e.g., she correctly believes that she has most reason to go to the doctor today, even if she is mistaken about the nature of this reason). Moreover, she does not suffer from weakness of will. When she comes to a settled judgment about what she has sufficient reason to do, she does it.

If RE is correct then from a practical point of view Deludia's behavior is beyond reproach, for what she does is endorsed by reasons. Consequently, there is nothing for which she as an agent might fairly be criticized⁶⁴ and nothing that can provide others with grounds for wanting to intervene in her life in an effort to help her engage with the reasons that, unbeknownst to her, really do best support her behavior. With respect to Deludia's visits to the doctor what matters on the RE view is that Deludia sees her doctor when she is in need of medical care, something she always does. That she fails to do so in light of the reasons that support her doing so does not matter. Reasons can endorse behavior that is not motivated by the reasons that endorse it.

Because Deludia's behavior is endorsed by reasons RE cannot explain what it is about her behavior that renders it defective. An appeal to RARP can do better because according to that principle Deludia does not do what she has sufficient reason to do, namely, to behave in reason supported ways *and* to do so in light of those reasons. When

⁶⁴ Because Deludia suffers from severe mental illness she is not blameworthy for behaving as she does and therefore the criticism to which she or her behavior are subject would not take the form of moral censure.

Deludia is ill she has reason to see a doctor because medical treatment is likely to promote and protect her health, a claim with which both RE and RARP are consistent. But when Deludia does not recognize and appropriately respond to this feature of a doctor's visit as a guide to her behavior she fails to do what she has sufficient reason to do, for one always has sufficient reason to make one's sufficient reasons action-guiding. RE is correct insofar as it is true that Deludia ought to go to the doctor when doing so will protect and promote her health. But RE is incomplete, for she ought to go to the doctor *because* it will promote and protect her health, not because it will cause her hair to turn red.

Deludia's case is analogous to Robert Nozick's well-known "experience machine" thought experiment.⁶⁵ Nozick asks us to imagine a machine that can stimulate our brains in such a way as to bring about any experience we might desire. Someone who has always wanted to climb Mt. Everest, or to have devoted friends, or to write a great book can have the experience of doing or having these things simply by allowing scientists to plug her into a sophisticated machine that can be programmed to deliver the appropriate experiences. Though the subjective states—that is, the experiences—caused by the machine are indistinguishable from the states that would be had were one actually to accomplish the things experienced, Nozick thinks we nonetheless would have good reason to reject the invitation to spend our lives plugged in to the machine. First, he argues that we want to do certain things and not just to have the experience of doing them and that moreover we want the experience of doing certain things only because we want

⁶⁵ Nozick, Robert, *Anarchy, State, and Utopia*, Basic Books, 1974, pp. 42-43

actually to do them.⁶⁶ Second, there is some kind of person that each of us wants to be—for example, we might want to be kind, or courageous, or accomplished—and a human blob plugged into a machine cannot be any of these things.⁶⁷ Third, Nozick claims that by plugging in to the machine we “limit ourselves to a man-made reality, to a world no deeper or more important than that which people can construct.”⁶⁸

Nozick's experience machine argument is meant to show that any moral theory that places primary value on people's having certain subjective mental states conflates the value of the subjective effect of doing or having something with the value of doing or having that thing. The idea is that if we would not choose to live our lives plugged in to the experience machine, then we must not care primarily about experiences. Nozick concludes that we have good reason to reject subjective state theories of well-being like hedonistic utilitarianism.

There are obvious differences between Deludia's detachment from reality and that of someone wired up to the experience machine. Deludia has not chosen which experiences she will have nor is anyone directly responsible for creating just the experiences she wants. Therefore, some of the concerns Nozick raises with respect to a life lived in the experience machine do not apply to Deludia's predicament. Deludia's life is not limited to a man-made reality, nor is she completely detached from the world in a way that in principle precludes the possibility of her being whatever kind of person she wants to be. But the most important difference between a life lived in the experience

⁶⁶ Ibid, p. 43

⁶⁷ Ibid

⁶⁸ Ibid

machine and Deludia's life is that Deludia does not merely have the experience of doing the things she has reason to do. Deludia really does behave in reason supported ways.

Nevertheless, her case is relevantly similar to that of someone plugged in to the experience machine insofar as neither of their “successes”—i.e., behaving in reason supported ways, having a pleasurable experience—is related in the right way to the features of the world that explain what it is about these things that make them worthy of having or doing. When the experience machine delivers the experience of winning the Nobel Prize the experience is not caused by winning the Nobel Prize. And Deludia's seeking medical care is similarly causally unrelated to the consideration that speaks in favor of her seeking medical care. Deludia's actions as well as the subjective states of the person in the experience machine are defective because they fail to correspond with the features of the world that govern behavior—in Deludia's case—or subjective states—in the case of the person in the experience machine. The subjective states of the person in the experience machine are not responsive to the states of affairs purportedly represented by the content of these states. Similarly, Deludia remains oblivious to the considerations that really do support her actions. What she takes to be her reasons are in fact figments of her imagination and what really are her reasons are beyond her capacity to recognize them. I noted above that some of Nozick's criticisms of a life lived in the experience machine do not apply to Deludia's situation because the cases differ in significant ways. However, a revised version of Nozick's first objection does apply. The original objection maintains that ultimately it is not experiences that people are after. Rather, what people want is actually to do, have, or be the things the purportedly desirable experiences

represent them as doing, having, or being. Deludia's case suggests an analogous conclusion, this time with respect to the role reasons play in our behavior.

Reflection on Deludia's situation reveals that a person's behavior can be defective even when it is supported by good reasons. More specifically, behavior is defective when these reasons fail to play a role in the processes that determine the behavior. Despite always doing what she has good reason to do Deludia is profoundly out of sync with the normatively significant features of her world, totally disengaged from the considerations that ought to govern her behavior. Insofar as we aspire to let our behavior be guided by reason we view Deludia's predicament with sorrow. Imagine a Reason-Endorsing Machine that could somehow guarantee that one always does what one has sufficient reason to do while leaving one completely disengaged from those reasons. One can imagine situations in which it would be better, on balance, to let one's life be governed by the Reason-Endorsing Machine. For example, it could very well be better to be influenced by the Reason-Endorsing Machine and consequently to behave in reason supported ways (albeit blindly) than to be both systematically detached from one's reasons *and* to behave in ways that are not endorsed by these reasons. But despite such possible benefits, those who are capable of recognizing and responding to reasons would not welcome the prospect of becoming entangled with such a machine.

Normatively Worthy Behavior

Manipulators intend to leave their manipulees detached from an important aspect of reality, namely, the reasons that ought to govern the manipulees' behavior. By failing to take seriously the role reasons ought to play in guiding behavior, manipulators aim to

cause their manipulees to exhibit behavior that is defective. In this section I characterize the nature of this sort of defect. In doing so, I draw on Julia Markovits's compelling account of what gives a morally right action *moral worth* and extend that account beyond morally right action and to behavior more generally.

Markovits wants to explain what it is about some actions that make them morally worthy. An action's moral worth as distinguished from its moral rightness was emphasized most famously by Kant, who argued that an action's rightness does not determine its moral worth. One of his examples is that of the merchant who does not overcharge an inexperienced customer. The merchant's customers are treated honestly but if the merchant behaves this way out of mere prudence, e.g., so that he does not develop a bad reputation which would lead to a reduction in his business, then the merchant's motives were selfish and hence not motivated by his recognition of his moral duty.⁶⁹ Markovits quotes Kant's radical claim that people "without any further motive of vanity or self-interest" who act in accordance with duty because they "find an inner pleasure in spreading joy around them and can rejoice in the satisfaction of others as their own work" perform acts that have "no true moral worth."⁷⁰ As Markovits points out, this is one of Kant's more controversial and unpopular claims, as it suggests that beneficent actions grounded in an agent's selfless desire to promote the happiness or well-being of others are entirely lacking in moral worth. This is so, according to Kant, because such actions are not performed from the motive of duty, i.e., the agent is moved not by her recognition that beneficence is a moral duty but rather by her simple desire to make

⁶⁹ Kant, Immanuel, *Grounding for the Metaphysics of Morals*, trans. James W. Ellington, Hackett Publishing Company, 1993, p. 10

⁷⁰ Markovits, Julia, "Acting for the Right Reasons," *Philosophical Review*, Vol. 119, No. 2, 2010, p. 202. The Kant quote can be found on page 11 of the Ellington translation of the *Groundwork*.

people happy/better off. Markovits calls this interpretation of Kant's account of moral worth the *Motive of Duty Thesis*.

Markovits wants to replace the *Motive of Duty Thesis*—that is, the thesis that an action has moral worth if and only if it is performed because it is right—with an alternative view that can vindicate Kant's emphasis on the role motives play in determining an action's moral worth without entailing some of the less attractive features of his view. Markovits labels her alternative account *The Coincident Reasons Thesis*, according to which

*my action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action—that is, if and only if I perform the action I morally ought to perform, for the (normative) reasons why it morally ought to be performed.*⁷¹

On this account, an agent who performs an action in light of the moral considerations that support that action performs an action with moral worth. It is neither necessary nor sufficient that the motive of the agent refers to the rightness of the action or to her duty to perform it. What matters is that the agent's behavior is guided by the moral considerations that *make* the action right. Markovits illustrates the distinction between the reasons that make an action right and the rightness of the action with an appeal to a scene from Mark Twain's *Huckleberry Finn*. Huck decides to protect his companion, the runaway slave Jim, rather than to turn him in to the slaveholders who are looking for him. He settles on this course of action despite his believing that in doing so he is stealing and therefore *doing wrong*.⁷² Huck is not motivated by the rightness of this action because he

⁷¹ Ibid, p. 205, emphasis in original

⁷² Ibid., p. 208. Markovits's Twain reference is to Twain, Mark, *The Adventures of Huckleberry Finn*, New York: Oxford University Press, 1996, pp. 271–72.

does not believe it is right. Instead, he is motivated by the considerations that make his action right, namely, that Jim seems to him to display common characteristics of humanity like love, decency, and friendship. According to the *Motive of Duty Thesis*, Huck acted rightly in saving Jim but his action lacked moral worth. According to the *Coincident Reasons Thesis*, Huck's action was both rightful and morally worthy. Clearly the latter view better accounts for our considered judgments about the moral worth of Huck's saving Jim.

Markovits's compelling account of an action's moral worth can help illuminate the wrongfulness of manipulation. When a manipulator intends a manipulee to behave in ways the manipulator believes to be supported by reasons and yet deliberately fails to make these reasons apparent as action-guides to the manipulee, the manipulator displays an indifference to what I will call the *normative worth* of the manipulee's behavior. Behavior has normative worth to the extent that it is motivated by the right reasons, that is, to the extent that the motivating reasons of the agent coincide with the reasons that justify the behavior. In understanding normative worth in this way, I am extending Markovits's account beyond right actions and to behavior more generally.

Markovits's account of the conditions that must obtain to render an action morally worthy bears a resemblance to the account of reason-tracking that occupies a central place in my account of manipulation. She too emphasizes how the relations between an agent's motivating reasons and her normative reasons can affect our evaluation of the character of the agent's behavior. However, the distinction she draws between actions with moral worth and those lacking it applies to contexts in which the only relevant reasons are those that justify actions *morally*. The question of normative worth, by

contrast, applies to any context in which it is appropriate to explain or justify behavior by an appeal to reasons more generally. The reasons in light of which manipulated agents act or those that justify their actions are not always moral reasons. This means that in trying to evaluate whether a particular action is the product of manipulation we will often have to consider reasons that lack moral salience. In other words, normative worth can attach to a much wider class of behaviors than moral worth can. As such, I understand behavior to have normative worth *if and only if the motivating reasons that explain the behavior coincide with the reasons that justify the behavior.*

Recall that according to RARP an agent who has sufficient reason to Φ has sufficient reason to Φ -for-that-reason. This principle, when conjoined with the account of normative worth just sketched, entails that the behavior of an agent who has sufficient reason to Φ and yet fails to Φ -for-that-reason lacks normative worth. On this view, Deludia's behavior is lacking in normative worth because the reasons that explain her behavior do not coincide with the reasons that support it.

Manipulators engaged in reasonable manipulation make use of means of influence that leave their manipulees detached from the considerations that ought to govern their behavior. Such behavior is as a result defective, lacking in normative worth.⁷³ Thus, RARP can explain why it is *pro tanto* morally impermissible deliberately to influence a reasons-responsive agent in a way that leaves her in this predicament. It can also help explain why Manipulation Principle 4 holds that it is not *pro tanto* impermissible to use

⁷³ One complication here arises with respect to manipulated behavior that does respond to good reasons that the manipulator supplies as a mere means to accomplish her end. A manipulator might cite good reasons because they would be more effective in motivating the desired behavior than bad reasons would be, though it is the bad reasons that motivate the manipulator. Ulterior motives often work in this way. Here the manipulator will not judge the manipulee's behavior to be lacking in normative worth. However, in such cases the normative worth of the action is of no consequence to the manipulator—she does not care one way or the other about the justificatory force of the reasons in light of which her manipulee behaves—and consequently expresses disregard for the person she seeks to influence.

this kind of influence on an agent who is not relevantly and decisively responsive to reasons, so long as the behavior being sought is endorsed by reasons. Because such agents are incapable of appropriately recognizing and responding to reasons the manipulator cannot be required to limit her modes of influence to just those that refer to justifying reasons. If an agent is incapable of non-defective action there can be no obligation on the part of those wishing to influence her to avoid causing her to behave defectively. In such cases it is sufficient from the moral point of view that the influencer aim to bring about behavior she believes to be endorsed by good reasons and that she believes *would* guide the behavior of the manipulee were the manipulee capable of recognizing and appropriately responding to these reasons.

As the case of Gabriel illustrates, the absence of responsiveness to reasons does not nullify these reasons but it does render an appeal to these reasons moot or even self-defeating. If Simon were to arrange for Gabriel to meet with the boss about a job under the pretense of introducing them at a party, his action would qualify as manipulative. However, because Gabriel's anxiety renders him relevantly and decisively unresponsive to the reasons that support his interviewing with Simon's boss, Simon's action may be justified, for in this case the benefit of employment may be significant enough so as to outweigh Gabriel's temporary detachment from the considerations that ought to guide his behavior.

Conclusion

If what I have said thus far is right then we have reason to avoid using means of influence that fail to track reasons. Such means of influence leave those whose behavior they target

detached from the considerations we believe ought to govern their behavior, consequently rendering their behavior lacking in normative worth.

It is important to stress the limits of the ethical analysis I have elaborated here. First, it does not tell us how to weigh the wrongness of manipulation against other considerations. People do not have absolute rights against being manipulated and so the question will arise in any given case of manipulation whether the wrongness of the influence is outweighed by any beneficial consequences that might accrue to the manipulee. We may also want to take into account the intentions of the manipulator, which in cases of paternalistic manipulation may be noble. *Manipulation Principle 4* is a “mid-level” ethical principle that can be evaluated in light of higher-level normative theories like consequentialism, Kantian deontology, virtue theory, and so on. The principle may provide a premise in an argument against some normative theories. Markovits argues that the *Coincident Reasons Thesis* provides us with an objection to utilitarianism because utilitarianism cannot offer as compelling an account of moral worth.⁷⁴ It may turn out that the analysis of manipulation I defend has similar consequences for utilitarianism or some other theory. I cannot explore this possibility here, though it does appear to me that utilitarians may struggle to explain why manipulation is wrong. In any case, the main point here is that more work needs to be done to incorporate my account of manipulation into broader discussions of philosophical ethics.

Second, the analysis is not meant to be comprehensive. In Chapter 1 I argued that manipulation does not necessarily involve wrongs like unjustified harm, violations of

⁷⁴ Markovits, pp. 230-237

autonomy, or deception. However, manipulation often does involve such things and when it does involve them manipulation is wrong for whatever reasons unjustified harm, violations of autonomy, or deception are wrong. What I have tried to show is that manipulation is *pro tanto* morally impermissible for the reasons elaborated above in addition to, and quite apart from, whatever else may be wrong with it. That is to say, even in cases where manipulation does not involve the common wrongs often associated with it, it is wrong because it is intended to leave its targets detached from an important aspect of reality, namely, the considerations that ought to guide their behavior.

Chapter IV

Manipulation and Libertarian Paternalism

In this chapter I first explain why an adequate account of manipulation is especially important in public policy, clinical, and medical research contexts. I then survey and criticize recent work on manipulation in the public policy and bioethics literature, focusing mostly on responses to Richard Thaler and Cass Sunstein's influential work on so-called "nudges." Worries about nudge-like influences are generally framed in terms of the *control* nudgers purportedly exert over their nudgees. I believe the emphasis on controlling influence is well-motivated, but as I will show it is not obvious how best to distinguish between controlling and noncontrolling influence. Thus, one of the main objectives of this chapter is to provide a more helpful notion of what controlling influence amounts to. On the view I defend, an influence is controlling when it is arbitrary, and an influence is arbitrary when the influence is not constrained by the reasons the influencer believes support the behavior she seeks to bring about. This notion of controlling influence nicely converges with my account of manipulation and helps remedy some of the confusion that pervades recent discussions of the ethics of nudging. My main conclusion is that nudges are manipulative and are therefore subject to the ethical analysis laid out in the previous chapter.

Thus far I have tried to provide a plausible general account of interpersonal manipulation and to explain what it is about this form of influence that makes it morally wrong (when it is). I now turn more specifically to manipulation in clinical medicine, medical research, and public health. Manipulation on the part of doctors, researchers, and

public health officials or governments seeking to influence citizens' health-related behavior is nothing new, and of course these agents do not have a monopoly on the use of manipulative forms of influence. Nevertheless, the public health, clinical, and research contexts present a number of distinctive considerations that make it imperative that we have an adequate analysis of manipulation and its normative significance.

The Ability to Manipulate: First, these are all areas in which there are significant power disparities between influencers and those they seek to influence. Physicians and medical researchers are viewed as authorities on questions of health and medical treatment, with many patients and subjects prepared to defer to their judgment.⁷⁵ However, very often—perhaps always—the considerations that count in favor of or against some possible course of action are grounded in a feature of the situation whose relevance is not strictly clinical—for example on a value, tradition, or set of preferences about which physicians or researchers cannot generally claim expertise. Even when patients or subjects are not prepared to defer, medical professionals are well-placed to employ their medical expertise and their general sophistication to influence the deliberations and decisions of their patients or subjects. Moreover, most of the time interactions between medical professionals and their patients or subjects take place in settings that make many non-professionals uncomfortable in the best of circumstances (for example the hospital or doctors' office). When non-professionals find themselves in such environments due to illness their capacity to resist the influence of a medical professional is compromised

⁷⁵ Frosch, Dominick, et al., "Authoritarian Physicians and Patients' Fear of Being Labeled 'Difficult' Among Key Obstacles to Shared Decision Making," *Health Affairs* (May 2012) vol. 31 no. 5 1030-1038, accessed March 20, 2013, doi: 10.1377/hlthaff.2011.0576.

even further. Consequently, patients and research subjects are vulnerable to an inordinate amount of influence from medical professionals.

Similarly, public health officials and governments enjoy substantial—often unparalleled—powers to influence citizens’ behavior. Legislators’ drafting and passing laws and the state’s monopoly on the use of force—the latter exercised through the courts and the police—are the most obvious examples of these powers. However, in addition to its coercive powers the state can influence behavior through various non-coercive means, such as tax breaks, educational campaigns, the creation of markets, the “bully pulpit”, and various programs implemented on the basis of incentives rather than the threat of force. Public health officials collect vast amounts of information about citizens’ health and their use of health care resources in order to understand where interventions are needed. They then go on to implement strategies intended to improve public health, with many of these strategies involving non-coercive forms of influence. With so much information and expertise, and often with the financial resources to implement their strategies, public health officials are in a good position to impact the behavior of millions of people.

The Temptation to Manipulate: People do always independently behave in ways that doctors, researchers, or public health officials want them to behave, and when it comes to medicine and public health people’s choices can have enormously important consequences. Consequently, medical professionals and public health officials sometimes will be tempted to manipulate their patients and subjects and public health officials will be moved to manipulate the behavior of citizens. For example, if a patient is hesitant to consent to a life-saving intervention, her doctor might reasonably judge that manipulating

her to provide consent is justified. Similarly, public health officials concerned about high obesity rates and the attendant social and economic costs might reason that the benefits of manipulatively “nudging” citizens toward healthier diets and more exercise is preferable to higher morbidity and mortality and out of control health care costs. With respect to research, investigators might be tempted to manipulate potential research subjects for a number of reasons. First, it is not always easy for investigators to find and enroll a sufficient number of subjects. Second, the hope that the research will lead to an important discovery that ultimately will be a great benefit to others can serve as a powerful motivator to conduct the study. Third, researchers can feel professional pressure to complete and publish their studies, for example when they know they will be competing for future funding or when they are seeking a promotion.

Advances in Behavioral and Social Scientific Research: Research in fields like cognitive science, psychology, and behavioral economics have allowed us better to understand human decision making and more accurately to predict the choices people make. With this growth in knowledge comes an improved ability to influence behavior, for to the extent that we understand decision-making processes and the relationship between these processes and the environment in which they take place we can not only predict how people will choose but also intervene in and influence their choices.⁷⁶ Richard Thaler and Cass Sunstein—whose work on “nudges” I will address in more detail below—are the

⁷⁶ Thus, the adoption of methods of influence derived from behavioral and social science can be understood as further improving the ability of government officials, policy experts, and health professionals to influence (and sometime manipulate) the behavior of those whose health-related choices are deemed in need of improvement. Nevertheless, I believe the recent advances in these sciences coupled with the growing interest in exploring their applicability to shape behavior justifies distinguishing these recent developments from the powers of influence these professionals have long enjoyed in virtue of their institutionally-granted authority, their social standing, or their specialized knowledge.

best-known proponents of adopting behavioral and social science research to promote decisions that can be advantageous on both the individual and social level.⁷⁷ These scientific advances and their possible value as policy instruments have not gone unnoticed by governments, policy makers, and health care professionals. The National Institutes of Health is committed to overcoming what it perceives as a “consistent difficulty in rapidly translating basic science discoveries into effective interventions”⁷⁸ and the NIH Office of Behavioral and Social Sciences Research and its partners have devoted resources to study and intervene in health-related behaviors.⁷⁹ Such interventions would be designed to reduce the amount of unhealthy behaviors (e.g., smoking, failing to exercise, etc.) and consequently to improve public health (e.g., with regard to cardiovascular disease, diabetes, and cancer).⁸⁰ Similar efforts are underway in other countries.⁸¹ Some recent and continuing research is more narrowly focused on influencing decisions made in clinical medicine. For example, a recent study sought to determining how the status quo or default bias (which causes people to favor the option that is the default option over alternatives) influences end-of-life decisions, such as the

⁷⁷ Sunstein, Cass and Richard Thaler, “Libertarian Paternalism is Not an Oxymoron”, *The University of Chicago Law Review*, Vol. 70, No. 4, (Autumn, 2003) and *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, 2008

⁷⁸ NIH Office of Behavioral and Social Science Research. Available at http://obssr.od.nih.gov/scientific_areas/health_behaviour/behaviour_changes/index.aspx (accessed March 20, 2013)

⁷⁹ Ibid.

⁸⁰ Ibid. The NIH provides a list of researchers who in 2010 received Common Fund awards to study the effects of various interventions on health-related behavior. The list can be accessed here: <https://commonfund.nih.gov/behaviorchange/overview.aspx> (accessed March 20, 2013).

⁸¹ In 2010 the Institute for Government and the Cabinet Office in the UK published a report entitled, “MINDSPACE: Influencing behavior change through public policy,” which showcases the ways in which research from the field of behavioral economics can be adopted by policy makers to address problems like crime, obesity, and environmental degradation. See <http://www.instituteforgovernment.org.uk/our-work/better-policy-making/mindspace-behavioural-economics> (accessed March 20, 2013).

decision whether to request comfort care rather than more clinically aggressive (and costly) medical interventions.⁸²

Of these three reasons why it is especially important in the medical and public policy contexts to get clear about manipulation, the latter has received the most attention, largely as a result of Thaler and Sunstein's advocacy of the use of "nudges." Their book⁸³ is written in an accessible style and the claims they advance there are in certain respects very attractive. For example, they maintain that nudges can leave individuals and society better off without placing any significant constraints on anyone's liberty. If they are correct, then their favored policies will have broad appeal across the political spectrum, for the standard liberty-based objections to interfering in people's lives will not apply.

Libertarian Paternalism and Nudges

Thaler and Sunstein (henceforth "T&S") can be understood as advancing two central claims. The first is a descriptive claim about the way people make decisions. The second is a normative claim about how knowledge about the way people make decisions should be used by policy makers and other "choice architects." A choice architect is the person responsible for structuring people's choice situations,⁸⁴ for example the person who decides how food should be presented in the cafeteria or who designs the forms

⁸² Halpern, Scott, et al., "Default options in advance directives influence how patients set goals for end-of-life care," *Health Affairs*, (February 2013) 32:2408-417, doi: 10.1377/hlthaff.2012.0895 (accessed March 20, 2013). This study is one of many conducted under the auspices of the Fostering Improvement in End-of-Life Decision Science (FIELDS) Program at the University of Pennsylvania's Perelman School of Medicine, which is partnered with universities, medical schools, private foundations, U.S. governmental agencies, as well as with health insurance companies. See <http://chibe.upenn.edu/fields-program> (accessed March 20, 2013).

⁸³ *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, 2008

⁸⁴ A "choice architect" is the person responsible for organizing people's choice situations. Thaler and Sunstein, *Nudge*, p. 3

employees fill out when they are making decisions about their retirement accounts. The descriptive claim is grounded in scientific research showing that human decision making is subject to a variety of tendencies that can lead to poor decisions. T&S describe a number of such tendencies, some of which are known as “cognitive biases” or “heuristics” in the psychological and behavioral economics literature. I have described some of these in earlier chapters, e.g., the “framing effect”—where the way information is framed can have significant effects on the choices people make, and the “status quo” or “default” bias—where people tend to choose the option that has been pre-selected for them rather than the option that requires them to “opt-out.”⁸⁵ Here are two more that are particularly striking and which would lend themselves rather easily to libertarian paternalistic purposes:

Anchoring: the piece of information with one has in mind as one begins deliberating can have an exaggerated impact on subsequent judgments. T&S discuss people who are asked to guess the population of Milwaukee, Wisconsin. People from Chicago use their city’s population as the “anchor” and adjust down, while people from Green Bay use the population of that city as their reference point and adjust up. Consequently, people from Chicago estimate Milwaukee’s population to be higher than do people from Green Bay.⁸⁶

Availability: people’s estimates of probabilities of events are strongly influenced by how easy it is for them to think of examples of events of the relevant type. For example, because terrorism gets much media attention and because terroristic events make such a

⁸⁵ Chapter 1

⁸⁶ Thaler and Sunstein, *Nudge*, p. 23

strong psychological impact, people overestimate the risk of being harmed in a terrorist attack. Conversely, they underestimate the risk of someone's dying from an asthma attack—examples of asthma attacks simply do not “spring to mind” in the way examples of terrorist attacks do.⁸⁷

The social scientific research T&S describe is of course quite interesting and important in its own right, but it is their recommendation for how this research should be used that has attracted attention from outside the scientific community and which raises difficult ethical questions. Their central claim is that our scientific knowledge of human decision-making processes should be put to use in efforts to influence people in ways that will lead to their making decisions that are better for them or for society. Rather than encroaching on their liberty by coercing people to do what is in their or society's interests, T&S argue that we ought instead to *nudge* them to do so. T&S define a ‘nudge’ as “any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives.”⁸⁸ Because nudges neither limit the set of options among which people choose nor attach heavy burdens to some of those options, T&S hold that nudges are *libertarian*.⁸⁹ They go on to argue in favor of a subclass of nudges, namely, those aimed at improving people's lives as judged by the people who are influenced.⁹⁰ Thus, for T&S a nudge is both

⁸⁷ Ibid., pp. 24-26

⁸⁸ *Nudge*, p. 6

⁸⁹ Ibid., pp. 4-5

⁹⁰ In other words, there is no conceptual connection between nudges and good choices. A person can be nudged into choosing badly or choosing well. T&S do not discuss nudges that are aimed at making people worse off, and thus it might be tempting to think that nudges always aim at improving people's decision-making. This would be a mistake.

libertarian and paternalistic when it preserves a person's freedom of choice while improving that person's life according to standards endorsed by the influenced person.⁹¹

It is worth noting that as it stands the definition of 'nudge' T&S provide is too broad to be of much help in defining the particular class of behavioral interventions that really interest behavioral economists or policy makers. If a nudge really is *any* aspect of the choice situation that preserves freedom of choice then just about any form of non-coercive influence—including straightforward rational persuasion—will count as a nudge. For example, when a physician informs a sick patient seeking a cure that there is a cheap or free drug that will cure his illness, the physician alters the patient's choice situation in a way that will affect the patient's behavior in a predictable way. However, T&S do not include cases of straightforward information provision or rational persuasion in their discussion of nudges. Nor do they include cases of outright deception, even though deception does not limit choice and can lead to improvements in the lives of the deceived.

T&S are interested in influences that not coercive or deceptive and which differ from straightforward rational persuasion. This is clear from the examples they use throughout their work, from their emphasis on particular kinds of imperfections in human decision making, and their short discussion of dual process theory. Regarding the latter, T&S describe "two systems" of cognitive processing. The "Automatic System" is fast, unconscious, uncontrolled, and effortless. The "Reflective System" involves processes that are conscious, controlled, and which require effort.⁹² The policy recommendations T&S advance involve using knowledge of how the Automatic System works in an effort

⁹¹ Thaler and Sunstein, *Nudge*, p. 5.

⁹² Thaler and Sunstein, *Nudge*, pp. 19-20

to steer people's decisions in directions that will improve their own lives as well as the lives of others. T&S write that "if people can rely on their Automatic Systems without getting into terrible trouble, their lives should be easier, better, and longer."⁹³ A nudge, then, is an alteration in the choice situation that predictably alters behavior via cognitive processes that are fast, uncontrolled, effortless, and unconscious. A nudge is both paternalistic and libertarian when it preserves the influenced person's freedom of choice while aiming at improving her life.

III. Objections to nudges

Critics have pointed out that some of the policies T&S propose may involve substantial constraints on choice and thus do not satisfy the conditions that distinguish nudges from coercion, a form of influence that is uncontroversially problematic from a moral point of view. For example, publicizing an "environmental blacklist" of companies that pollute⁹⁴ may amount to placing significant costs on some options the company may be inclined to choose, namely, the option to continue (or to start) conducting business in a manner that causes pollution. If these costs (measured in terms of social opprobrium and its attendant effects on a company's bottom line) would be substantial then the threat of being blacklisted would make it very difficult or even impossible for companies to avoid choosing the course of action favored by the choice architects. Clearly this would not respect these companies' freedom of choice in the sense of "freedom of choice" T&S stipulate their policies are meant to respect.⁹⁵

⁹³ Ibid., p. 22

⁹⁴ Ibid., pp. 190-193

⁹⁵ Hausman, Daniel and Brynn Welch, "Debate: To Nudge or Not to Nudge," *The Journal of Political Philosophy*, Vol. 18, No. 1 (2010), p. 125

In response to this objection, T&S might concede that some of their proposals do not sufficiently preserve freedom of choice and thus do not qualify as libertarian paternalist nudges. This does not mean that such policies cannot be justified, but it does mean that whatever justifications are offered will be unlikely to satisfy the libertarian, as the considerations counting in favor of the policies would have to be weighed against their encroachments on liberty. For example, in the pollution case the limitations a blacklist would place on a company's liberty to conduct business any way it sees fit would have to be weighed against the public health and environmental benefits brought about by a reduction in pollution. In any case, putting aside concerns directed at policies that may not, contrary to T&S's initial claims, respect freedom of choice, there are independent worries about the kind of influences that do preserve freedom of choice as that freedom is conceived according to T&S's account of libertarian paternalism.

The growing scholarly literature on the ethics of nudging reveals that much of the concern about this form of influence is grounded in the view that nudges are controlling or threatening to the autonomy of those who are nudged. Jennifer Blumenthal-Barby and Hadley Burroughs write, "...one should consider whether the use [of nudges] would count as an instance of manipulation, as manipulation always involves some infringement on a person's autonomy."⁹⁶ On their view, manipulation infringes on autonomy "by virtue of it bypassing a person's capacity for reason."⁹⁷ According to Daniel Hausman and Brynn Welch, nudges threaten autonomy insofar as they undermine the influencee's "control over [her] own evaluations and deliberation," which control Hausman and

⁹⁶ Blumenthal-Barby, Jennifer and Hadley Burroughs, "Seeking Better Health Care Outcomes: The Ethics of Using the 'Nudge,'" *The American Journal of Bioethics*, 12:2, pp. 1-10. Blumenthal-Barby and Burroughs also cite other considerations they believe are ethically relevant. My focus here is on manipulation, and thus I will not address these other considerations here.

⁹⁷ *Ibid.*, p. 5

Welch identify with autonomy.⁹⁸ Yashar Saghai also makes an agent's control over her choices central to his analysis of nudging.⁹⁹ Though he does not think the term 'autonomy' is a helpful one in this context,¹⁰⁰ the considerations that on Saghai's view matter with respect to our moral evaluation are similar to, if not identical with, the considerations that are relevant to our evaluation of an agent's autonomy: "the degree to which others control our choices and engage our deliberative capacities."¹⁰¹ Drawing on the work of Joseph Raz—who also argues that manipulation is wrong (when it is wrong) because it undermines autonomy¹⁰²—T.M. Wilkinson maintains that "[w]hat is primarily wrong with manipulation is that it violates autonomy"¹⁰³ and therefore that nudges are wrong when they are manipulative. Luc Bovens worries that, "[t]here is something less than fully autonomous about the patterns of decision-making that *Nudge* taps into. When we are subject to the mechanisms that are studied in 'the science of choice', then we are not fully in control of our actions."¹⁰⁴

All of these criticisms are grounded in the same concern over the extent to which nudges interfere with or undermine an agent's control over her evaluations, deliberations, and actions. Also, in their discussions of nudge-like manipulative interventions each of

⁹⁸ Hausman and Welch., p. 128. Hausman and Welch opt to use the term "shaping" rather than "manipulation" because they worry that the negative connotations carried by the latter term may make it appear that they are begging the question against defenders of nudging (p. 128-129). The difference between "shaping" and "manipulating" behavior seems on their view to be purely rhetorical and thus I read their claims regarding the ethics of shaping behavior as claims about manipulating behavior.

⁹⁹ Saghai, Yashar, "Salvaging the Concept of Nudge," *Journal of Medical Ethics*, Published Online First: February 20, 2013, doi:10.1136/medethics-2012-100727

¹⁰⁰ Ibid., p. 3

¹⁰¹ Ibid., pp. 6-7

¹⁰² Raz, Joseph, *The Morality of Freedom*, Oxford: Clarendon Press, 1986, p. 378, p. 420

¹⁰³ Wilkinson, T.M., "Nudging and Manipulation," *Political Studies*, article first published online September 7, 2012, doi: 10.1111/j.1467-9248.2012.00974.x

¹⁰⁴ Bovens, Luc, "The Ethics of *Nudge*," in Till Grune-Yanoff and S.O. Hansson (eds.), *Preference Change: Approaches from Philosophy, Economics and Psychology*, Theory and decision library A (42), Springer, 2009, pp. 207-219.

the authors quoted above is careful to distinguish nudges from rational persuasion,¹⁰⁵ with the possible exception of Wilkinson.¹⁰⁶ It seems that for these authors, that a form of influence differs from rational persuasion strongly suggests that it fails to respect an agent's control over her behavior, while rational persuasion does respect an agent's control over her behavior. This is an intuitively attractive idea but it is not at all clear that the ethical analysis of manipulation should depend on the distinction between rational persuasion and manipulation. This is because rational persuasion is sometimes manipulative.¹⁰⁷ Nor is it clear that emphasizing an agent's control over her behavior—as the authors just surveyed do—can help adequately distinguish rational persuasion from nudges or other morally problematic forms of manipulation. I now turn to a discussion of the relations between manipulation, control, and rational persuasion.

Manipulation, Persuasion, and Control

A good place to start is with Tom Beauchamp and James Childress's popular textbook, *Principles of Biomedical Ethics*, in which the authors distinguish between manipulation, coercion, and rational persuasion.¹⁰⁸ Though Beauchamp and Childress's discussion of manipulation is not limited specifically to nudge-like influences they do cite the framing

¹⁰⁵ Hausman and Welch, p. 128, Blumenthal-Barby and Burroughs, p. 5, Saghai, pp. 4-5, Bovens, pp. 3-5 (Bovens does not use the term "rational persuasion", though it seems clear that this is what he has in mind when he contrasts nudges from other forms of influence in which the content of the information provided by the influencer is what does the work.)

¹⁰⁶ On page 3 of his essay, Wilkinson expresses doubt about whether there is a clear conceptual distinction between rational persuasion and "supposedly irrational or non-rational methods." However, in relying on Raz's definition of manipulation to account for the wrongness of nudges, Wilkinson may be committing himself to such a distinction. This is because according to Raz manipulation "perverts the way that person reaches decisions, forms preferences or adopts goals." See Raz, pp. 377-378

¹⁰⁷ Chapter 1, pp. 32-51

¹⁰⁸ Beauchamp, Tom and James Childress, *Principles of Biomedical Ethics*, 6th ed., Oxford University Press (2009)

effect as one example of the sort of influence they consider to be manipulative.¹⁰⁹ And as we will see, the framework they provide for distinguishing different forms of influence would clearly count nudges as manipulative.

Beauchamp and Childress characterize manipulation by reference to other forms of influence, namely coercion and persuasion. They conceive of coercion as a form of influence in which one person controls another through the intentional use of “a credible and severe threat of harm or force.”¹¹⁰ Persuasion is a form of influence in which a person “come[s] to believe in something through the merit of reasons another person advances.”¹¹¹ Finally, manipulation is “a generic term for several forms of influence that are neither persuasive nor coercive. The essence of manipulation is swaying people to do what the manipulator wants by means other than coercion or persuasion.”¹¹² According to Beauchamp and Childress when manipulation is morally problematic it is because it is “incompatible with autonomous choice,” which incompatibility they seem to diagnose in terms of the “controlling influence” exerted on the manipulee by the manipulator.¹¹³

As it stands this schematic account fails to distinguish coercion from persuasion in a way that allows for a helpful characterization of manipulation as that form of influence that is neither coercive nor persuasive. This is because not only persuasion but effective coercion involves the provision of reasons—sometimes very good reasons. For example, that one will be shot if one fails to surrender one’s money is a very good reason to surrender one’s money. Perhaps the idea is that coercers provide coercees with a

¹⁰⁹ Ibid., p. 134

¹¹⁰ Ibid, p. 133

¹¹¹ Ibid.

¹¹² Ibid. p. 133-134

¹¹³ Ibid., p. 134 It is important to note that for Beauchamp and Childress controlling influence—influence that violates autonomy—is sometimes justified.

particular *kind* of reason, more specifically a reason having to do with force or serious harm, which force or harm it is in the power of the coercer to exercise or bring about. This seems right, but nevertheless it is false that coercion does not involve the provision of reasons while persuasion does. We might try to preserve the Beauchamp and Childress account by putting aside differences in the kinds of reasons coercers and persuaders offer and focusing instead on the more general attribute these forms of influence share, namely, the provision of reasons. According to this modified view, coercion and persuasion can be distinguished from manipulation insofar as only the former two forms of influence involve the provision of reasons, albeit reasons that differ in kind from one another.

This modified version of Beauchamp and Childress's view will not work either. The problem is that a manipulator may exploit her knowledge of a manipulee's set of propositional attitudes to sway the manipulee to do what she, the manipulator, intends that he do, and, crucially, the manipulee's behavior may be supported by what he, the manipulee, regards as good reasons. Take the following case:

Painkillers After Surgery: A doctor believes her patient has decisive reason to consent to a surgery. The patient is very reluctant to do so and his reluctance stems from what the doctor regards as a silly superstition. The patient is totally unresponsive to the reason his doctor believes supports his consenting to the surgery. He comprehends what his doctor tells him but simply does not regard it as compelling, given his beliefs, desires, values, and so on. But imagine that this patient is a recreational user of prescription analgesics, and that, knowing this, in her next meeting with the patient the doctor deliberately emphasizes the fact that these drugs are as a matter of course prescribed post-operatively. The doctor does not regard the provision of analgesics to constitute a good reason to consent to the surgery, but she

does recognize that her patient regards it as such. As intended, the patient consents to the surgery and does so in response to what he regards as a good reason, namely, that doing so will allow him legally to gain access to prescription painkillers.¹¹⁴

This case involves no coercion—there are no threats of force or harm. It does seem to qualify as a case of persuasion—and therefore not one of manipulation—on Beauchamp and Childress’s account, as the patient’s behavior is determined by his recognition of a reason, which reason has been advanced by his doctor. Moreover, the doctor did not subvert her patient’s autonomy, as the patient made his choice after deliberating on the basis of preferences and values he reflectively endorses. But nevertheless I think it is fair to characterize the doctor’s behavior as manipulative and to judge that the patient’s choice to consent to the surgery has been manipulated. The doctor succeeded in manipulating her patient *by* rationally persuading him and she did so without violating his autonomy. If this is right, then Beauchamp and Childress’s schematic account fails to distinguish manipulation from persuasion and incorrectly analyses the wrongness of manipulation exclusively by reference to autonomy violations.

In their essay on the use of manipulation in the recruitment of research subjects, Amulya Mandava and Joseph Millum distinguish manipulation from persuasion, claiming that the latter by definition violates autonomy while the former does not.¹¹⁵ Their understanding of persuasion is consistent with Beauchamp and Childress’s:

¹¹⁴ To avoid worries about the bypassing of the patient’s rational capacities, we can stipulate that he is not addicted to these drugs but merely enjoys taking them occasionally as part of a lifestyle he reflectively endorses.

¹¹⁵ Mandava, Amulya and Joseph Millum, “Manipulation in the Enrollment of Research Participants,” *Hastings Center Report* 43, no. 2 (2013), p. 38-47. DOI: 10.1002/hast.144

One agent persuades another to pursue a specific action when she motivates him by showing rational links between his existing set of reasons to act and that action. She can do this by showing logical connections between his existing reasons and the act she wants him to perform, or by honestly presenting facts that are relevant to his reasons to act.¹¹⁶

According to Mandava and Millum, persuasion “does not illegitimately interfere with” the influencee’s decision-making process and thus is a form of influence that—unlike manipulation—is “respectful of autonomy.”¹¹⁷ But on their account of persuasion a researcher who emphasizes the provision of analgesics when seeking the consent of a reluctant potential participant (for example by saying to him, “You would also receive painkillers, something I know you like”) persuades him to consent, as the researcher here “show[s] logical connections between [the subject’s] existing reasons and the act she wants him to perform.” This influence does not interfere with the participant’s decision-making processes and thus respects his autonomy,¹¹⁸ which entails (for Mandava and Millum) that the doctor does not manipulate her patient. But clearly the doctor’s behavior here is problematic and seems intuitively to be manipulative.¹¹⁹

Even if one has doubts about whether the painkiller cases qualify as instances of manipulation it should be clear that there is a salient difference between them and cases in which a person responds not only to what he regards as a good reason, but to considerations that both he *and the person influencing* him believe support his behavior. Thus, even if we think the doctor/researcher’s behavior can be justified, e.g. in light of

¹¹⁶ Ibid., p. 39

¹¹⁷ Ibid.

¹¹⁸ I stipulate that 1. the doctor does not believe the provision of the painkillers provides the potential research participant with a good reason to consent and 2. the participant is not addicted to painkillers and takes them recreationally as part of a lifestyle he reflectively endorses.

¹¹⁹ Even those who do not share the intuition that the doctor manipulates his patient can acknowledge that her emphasis on the painkillers requires justification in a way that her emphasis on the clinical benefits of the surgery does not.

the benefit the patient/subject or society would gain as a result of his consenting, it is plausible that her targeting of her patient/subject's penchant for prescription painkillers is something to which she is entitled to resort only after another means of influence—namely, the provision of what she, the doctor/researcher, takes to be good reasons—have failed. I take it that it is *pro tanto* impermissible for her to focus on the painkillers without first trying to influence her patient/subject by describing those features of the surgery/research she believes really do speak in favor of his consenting to it—e.g., that it will extend his life, decrease his level of discomfort, contribute important knowledge to the medical field, or whatever.

Persuasion and Control

In addition to their serving as counterexamples to the claims that 1. manipulation is necessarily distinct from rational persuasion and 2. manipulation necessarily violates autonomy, the painkillers cases also lead to related questions about the role the notion of 'control' can play in distinguishing morally problematic influence from morally benign influence. These questions are relevant to the debate over the use of nudges and manipulation more generally, as most of the authors surveyed above take an agent's lack of control over her evaluations, deliberation, and action to be the defining characteristic of behavior that, when intentionally caused by another agent, renders the influence morally problematic.

We typically do not consider rational persuasion to be a controlling form of influence. Yet, it is unclear how to specify a purely descriptive notion of 'controlling influence' in a way that distinguishes typical, morally unproblematic cases of rational

persuasion from coercion and manipulation. I will suggest below that my account of manipulation can explain the difference between controlling influence and non-controlling influence and that by doing so can also distinguish between “normal” non-manipulative rational persuasion and morally problematic manipulative persuasion. Before turning to my account I want to explain in more detail what I think is wrong with some other accounts of influence that emphasize an agent’s control over her behavior.

Beauchamp and Childress do not provide an account of what distinguishes controlling influence from non-controlling influence, claiming that “[i]n biomedical ethics we need only establish general criteria for the point at which influence threatens autonomous choice, while recognizing that in many cases no sharp boundary separates controlling and noncontrolling influences.”¹²⁰ This lack of clear guidelines to help distinguish controlling influence from noncontrolling influence is problematic, for Beauchamp and Childress analyze autonomous choice in terms of “normal choosers who act (1) intentionally, (2) with understanding, and (3) without controlling influences that determine their action.”¹²¹ Thus, they characterize autonomous choice in part by reference to the lack of controlling influence while later claiming that in many cases “no sharp boundary” exists between controlling and noncontrolling influence. But in the absence of criteria to distinguish controlling from non-controlling influence, we will have a difficult time ascertaining the point at which influence threatens autonomous choice. If the notion of control is to play a central role in the analysis of nudges or manipulation more generally, it needs to be specified in more detail.

¹²⁰ Beauchamp and Childress, p. 134.

¹²¹ Ibid., p. 101

The articulation of a more useful notion of control plays a central role in Saghai's essay, the overall aim of which is to provide an account of nudges that is free of the ambiguities with which the term was originally introduced by Thaler and Sunstein. Saghai offers an analysis of control intended to help distinguish between influences that preserve meaningful freedom of choice—as nudges are supposed to do—and those that do not. Early in his paper, Saghai claims that “[i]nfluences can be situated on a continuum from fully controlling to fully noncontrolling” with coercion being “fully controlling,” persuasion being “never controlling,” and with non-coercive, non-persuasive methods of influence making up a third category, which is itself divided into two sub-categories, “substantially controlling” and “substantially noncontrolling.”¹²² In order for an influence to qualify as a nudge it must be (*inter alia*) substantially noncontrolling, i.e., it must preserve the influenced agent's freedom of choice (so that the influence respects libertarian constraints).

Saghai's strategy is to provide a general account of non-persuasive, choice-preserving influence by reference to his substantial control/noncontrol distinction and then to define nudges as those influences that are (*inter alia*) substantially noncontrolling. His first step is to define *substantially noncontrolling influence*: “The Substantial Noncontrol Condition. A's influence to get B to ϕ is substantially noncontrolling when B could easily not ϕ if she did not want to ϕ .”¹²³ Next, he spells out what makes an influence easily resistible: “A's influence is easily resistible if B is able to effortlessly oppose the pressure to get her to ϕ if she does not want to ϕ .”¹²⁴ Analyzing “easily

¹²² Saghai, p. 2. Thus, on Saghai's account the form of influence that Beauchamp and Childress label “manipulation” can be further analyzed as either substantially controlling or substantially noncontrolling.

¹²³ Ibid.

¹²⁴ Ibid., p. 3

resistible” in terms of effortless opposition does not take us very far, but Saghai provides a more detailed account of the ability easily to resist an influence:

B is able to easily resist A’s influence when:

1. B has the capacity to become aware of A’s pressure to get her to ϕ (attention-bringing capacities);
2. B has the capacity to inhibit her triggered propensity to ϕ (inhibitory capacities);
3. B is not subject to an influence, or put in circumstances that would significantly undermine the relatively effortless exercise of attention-bringing and inhibitory capacities.¹²⁵

An influence is substantially non-controlling, then, when it is easy for an agent to resist it, and it is easy for her to resist the influence when she can recognize it and inhibit its effects on her behavior.

There are at least two problems with this analysis of substantially noncontrolling influence. The first is that it is too strict. It is too strict because it requires that an agent have the capacity to recognize and engage with the influence. Intuitively B easily resists A’s influence when it takes little or no effort for B to avoid behaving as A intends that she behave. B can do this even when she is unaware of A’s pressure. Suppose B is very depressed and thus does not recognize that A is trying subtly to seduce her. It is easy for her to resist A’s influence because she does not recognize it. Or B may not be disposed to ϕ to begin with and thus there is no propensity for A to trigger and for B to inhibit. For example, suppose B simply lacks the propensity to choose food that is placed at the front of the cafeteria line. In fact, she is strongly disposed always to select food from the end of the line. Intuitively, when A places the fruit and vegetables at the front of the line in an

¹²⁵ Ibid.

effort to get B to choose them, it is easy for B to resist his influence. A has still influenced B's choice—B chooses the junk food A placed at the end of the line—but nevertheless it was easy for B to resist A's influencing her to choose the healthy option.

The second problem with Sagha's analysis of substantially controlling influence is that it is too permissive. It is too permissive because it does not exclude garden-variety rational persuasion. To the extent that one is rational one will find it difficult to inhibit one's propensity to behave in accordance with the dictates of reason. For example, suppose A wants B to choose a generic drug rather than the name-brand equivalent. A knows that B cares very much about controlling health care costs and also that B is extremely critical of pharmaceutical company X. A informs B that the drug B takes, which is manufactured by company X, is available in generic form. A explains that the generic drug is much less expensive than the name-brand drug and that it is manufactured by some firm of which B is less critical than she is of company X. As a rational agent B has a propensity to act in accordance with her recognition of what she has most reason to do. B's beliefs, values, and preferences make it extremely difficult for B to continue choosing the name-brand drug. Moreover, B has "the capacity to become aware of A's pressure to get her to" switch to the generic drug, i.e., she sees very clearly that A is trying to get her to switch from the name-brand drug to the generic. On Saghai's view, A's influence is substantially controlling with respect to B's decision to switch from the name-brand drug to the generic. Yet, A merely provides B with information that, when combined with B's background attitudes, values and commitments, makes it rational for B to do what A wants her to do.

I have already noted that Saghai claims that rational persuasion is “never controlling,”¹²⁶ an assertion I have just argued is at odds with his analysis of substantially controlling influence, on which rational persuasion does seem to qualify as substantially controlling. Saghai might respond to this objection by amending his account of substantially controlling influence, e.g., by pointing to some feature of rational persuasion that disqualifies it from inclusion as a substantially controlling influence. Indeed, in initially characterizing persuasion as a form of influence that is “never controlling” Saghai remarks that this is because “the persuadee willingly accepts the reasons she is given.”¹²⁷ Saghai makes this parenthetical comment in the context of distinguishing coercion from persuasion and thus presumably the idea is that *unlike coercion* persuasion involves a willingness on the part of the influenced party to accept the reasons that are advanced by the agent who is seeking to influence her. This idea has some intuitive force, for there is a sense in which coerced actions are done against the will of the coercee. However, it is not obvious what to make of this, as there is also a clear sense in which coercees do willingly accept the reasons advanced by their coercers, e.g., when facing a gun they accept that they will be better off if they surrender their money than they would be if they do not.

So, both coercees and persuadees respond to reasons. Neither may be capable of inhibiting their propensity to do what the influencer intends that they do. Perhaps the notion of “willingness” at play in Saghai’s account can be vindicated by an appeal to the desires of the influencee: the mugging victim does not want to surrender his wallet, while someone who has been persuaded into donating money to a charity does want to make a

¹²⁶ Saghai makes this claim on page 2.

¹²⁷ Ibid.

donation. There are problems with this proposal as well. It is not true that we always want to do what others successfully persuade us we ought to do. One might donate to a charity but only grudgingly, against one's strong inclination to keep the money.¹²⁸ "Willingly accepting reasons" is not equivalent to "happily accepting reasons." In any case, if we always behave on the basis of our desires alone, then the coercee too does what she wants to do, i.e., she wants to hand over her money because she wants to avoid the extremely negative consequences attaching to her failure to do so.

Control and Arbitrariness

The crucial difference between coercion and persuasion is that only coercion involves a threat of serious harm, which harm it is in the power of the coercer to bring about. Any meaningful distinction between controlling influence and noncontrolling influence must recognize this difference. I propose that coercion is a controlling form of influence while persuasion is not because persuaders are constrained by the reasons they believe do (or do not) support the behavior at which they aim, which is to say that (typically) rational persuasion is a process of influence that tracks reasons. Coercers, on the other hand, are not constrained by reasons they believe support the behavior they seek—they use threats to generate new reasons.¹²⁹ Consequently, the will of the influencee is more closely tied to that of the influencer in cases of coercion than it is in cases of rational persuasion. For example, that someone is putting a gun to my head gives me an "open-ended" reason to do what he tells me to do, *whatever that might be*. Coercion works by subjecting the will of the coercee to the will of the coercer, whatever the content of the coercer's will. Such

¹²⁸ Kant in the *Groundwork* discusses such cases. Get citation if needed for more detail.

¹²⁹ Cite Chapter 1, where I discuss difference between manipulation and coercion.

influence is *arbitrary*, as the will of the coercer is responsive to the will of the coerced rather than to independent considerations. The arbitrariness of coercive influence is what makes it controlling.

The coercer who holds a gun to my head provides me with a reason to surrender my wallet, or to paint a house, or to hum a Beethoven symphony—in short, to do whatever he demands of me. By contrast, persuaders do not offer reasons that are open-ended in this way. A reason to believe that P is not applicable to the belief that Q (unless, of course, there is an appropriate rational connection between P and Q). Persuaders advance reasons that support particular beliefs or courses of action, reasons with no applicability to other beliefs or courses of action. Though rational persuasion can be very difficult to resist—sometimes perhaps just as hard to resist as coercion is—the pressure on the will of the persuadee is *not* arbitrary. Its source is the normative force of reasons that are themselves independent of the will of the influencer, even when what these reasons prescribe is identical to what the influencer wills.

I have just argued that coercion is controlling because it is an arbitrary form of influence and I suggested that an influence is arbitrary when it is not constrained by reasons the influencer believes support the behavior she intends to bring about. This way of construing ‘controlling influence’ gets the right result with respect to rational persuasion and coercion: (typical) rational persuasion is not controlling while coercion is controlling.

According to the account of manipulation I have been defending, manipulative influence is controlling because it amounts to arbitrary influence. Unreasonable manipulation—where manipulators aim at behavior they do not believe to be supported

by reasons—is controlling. When a manipulator intends a manipulee to behave in ways the manipulator does not believe to be supported by reasons (from the perspective of the manipulee) then clearly the influence is not constrained by reasons the influencer believes *do* support the behavior she intends to bring about. Paternalistic Reasonable Manipulation is also controlling. Paternalistic manipulation aims at reason-supported behavior but makes use of means of influence that are “open ended”, i.e., they involve normatively irrelevant features of the choice situation and thus can lead to reason supported behavior *or* behavior that is unsupported by reasons, depending on the whim of the influencer. For example, the same cognitive bias that that leads a patient to consent to a life-saving surgery could have been employed in a way that would have lead the same patient in the same circumstances to refuse the surgery. Non-paternalistic reasonable manipulation is also controlling. Recall that non-paternalistic reasonable manipulation happens when the manipulator believes the behavior she is seeking is supported by reasons but is not motivated by these reasons. Such influence is arbitrary because the supporting reasons do not constrain the manipulator’s intentions. For example, a manipulator who believes the only way he will succeed in getting a manipulee to do what he wants her to do is by citing the reasons he (the manipulator) believes do support the behavior and yet who remains motivated by some other considerations that he does not think support the behavior (i.e., he acts on “ulterior motives”), engages in influence that is arbitrary because the supporting reasons do not play a role in the influencer’s end. In such cases the supporting reasons are incidental to the manipulator’s aim—at bottom it is his will that determines the behavior of his influencee.

Are nudges controlling?

Saghai argues that nudges are best understood as substantially *noncontrolling* influences because they are easy to resist.¹³⁰ However, I have just argued that easy resistibility cannot be a necessary condition for noncontrollingness, for if it were rational persuasion would sometimes qualify as a controlling influence. I have also suggested that some influences that would count as easily resistible on Saghai's account—e.g., the use of framing effects—can be controlling because such influences are “open ended”, that is, they are grounded in the will of the influencer rather than in independent, normatively-relevant features of the choice situation. If an influencee does X rather than not-X because, and only because, the influencer wills that the influencee do X, then the influence leading to the doing of X is controlling. For example, the framing effect can be used to promote reason-supported behavior *or* reason-unsupported behavior, and which direction the influence takes is determined entirely by the will of the influencer rather than by the considerations that do (or do not) support the influenced behavior.

What about some of the other examples Thaler and Sunstein discuss? In *Cafeteria*, the choice architect arranges the food such that healthier items are chosen more often than they would be under some other arrangement (e.g., she places healthy food first in line).¹³¹ Clearly the fact that one encounters the salad five feet before one will encounter the French fries does not provide one with a reason to choose the salad rather than the French fries.¹³² If it were the case that placing the French fries first would

¹³⁰ Saghai, p. 5

¹³¹ Thaler and Sunstein, *Nudge*, pp.1-3

¹³² This claim assumes that choosing the items that come later would be easy to do. If the costs of failing to choose the items placed first in line are significant—e.g., one cannot see the options down the line and hence would be taking the risk that there is nothing that one would like to an equal or greater extent—then one will have good reason to select items that come first.

have led to increased consumption of these and reduced consumption of salad, then the influence exerted by the choice architect qualifies as open-ended—the choices people make are tied to the will of the architect rather than to the considerations that are directly relevant to their deliberations, e.g., that salad is healthier or that French fries are (arguably) more flavorful. Such influence is arbitrary and controlling.

What about taking advantage of people's propensity to "go with the flow" by making whatever option is favored by the choice architect the default option? Thaler and Sunstein discuss how employers can increase savings rates by automatically enrolling their employees into retirement plans while making it easy for employees to opt out if they choose.¹³³ Because the status quo bias does not track features of the choice situation that are normatively relevant, the exploitation of this bias also amounts to an open-ended form of influence. For example, an employer who for whatever reason wanted to decrease savings rates could make non-enrollment the default option. Thus, the behavior of employees with respect to their enrollment is tied not to the considerations that speak in favor of (or against) their enrollment, but rather to the will of the agent who structures the choice environment. The influence exerted by the choice architect is arbitrary and controlling.

There is an important qualification that should be amended to the preceding argument. It may be that in some cases a person who is influenced by a nudge behaves just as the nudger intended and yet is not nudged. Consider a person who chooses to consent to a medical procedure after having been informed that the survival rate for patients who undergo the procedure is 90%. The choice architect deliberately framed the

¹³³ Thaler and Sunstein, *Nudge*, pp. 108-109. They discuss the same example in "Libertarian Paternalism is Not an Oxymoron", pp. 1159-1160.

information in terms of the survival rate rather than in terms of the mortality rate in order to push the patient to consent, i.e., the choice architect intended to nudge the patient. But suppose the patient would have consented in any case, that is, even if the information had been framed in terms of the 10% of patients who do not survive. Though this patient was subject to influence that was arbitrary and hence which sought to control her behavior, she remained free of that control because the arbitrary device (the frame) played no causal role in her decision. Consider a more extreme case: someone who surrenders his cash to a mugger not because he is threatened—he is in no way moved by the threat—but because he recognizes that the mugger must be truly desperate to act in this way. When he hands over his money he does so because he sympathizes with the mugger and genuinely wants to help him. Here the mugger intends to control his mark, the mark does what the mugger intends that he do, and yet the mugger does not control the mark's behavior. Thus, the fact that a form of influence is controlling does not entail that the influenced behavior or the person exhibiting have been controlled.

In order to determine whether a piece of behavior is the product of controlling influence we need to know about the motives that generate the behavior. A default bias nudge might be effective because the influenced person simply “goes with the flow.” Such a person would have in those circumstances chosen whatever option was placed as the default. But if the person would have selected the same option even if it were not the default, then default bias played no causal role in her decision and hence did not control her behavior—even if the choice architect sought to control it. There are more difficult cases, too, such as those where a person “goes with the flow” but only because they are justified in believing that this is the best way for them to decide in these circumstances.

Consider an employee who knows her employer to care very much about the well-being of employees. She knows the leadership of the company consistently looks out for the interests of their employees and try, whenever they can, to do what is best for them. This employee might be justified in believing that if her employer made enrollment into a savings plan the default option there must be a good reason for her to go along with that choice. If such considerations lead her to select the default option, she has not been controlled by her employer, though a colleague of hers who makes the same choice would be controlled if his choosing the default option were *merely* the result of his “going with the flow.”

The Ethics of Nudging

Insofar as nudges are “open-ended” influences in that they are not constrained by the reasons that support the behavior at which they aim, they are manipulative.

Consequently, nudges are morally problematic for the reasons set forth in Chapter 3.

Nudges may be left detached from the considerations that ought to govern their behavior and to the extent that this is true their behavior is lacking in normative worth.

Consequently, it is *pro tanto* morally impermissible to nudge reasons-responsive agents.

However, this does not mean that it is always all-things-considered wrong to use nudges to influence people’s behavior. Nudges that are used as a “second line” method of influence where rational persuasion is impossible or otherwise inappropriate and which aim at reason-supported behavior can be justified. Nudges that are implemented as the preferred “first line” method of influence, or nudges that aim at behavior the nudger

knows to be questionable with respect to how well it is supported by reasons will be more problematic. The details of the specific case matter.

Consider T&S's discussion of automatic enrollment in retirement savings plans. They point out that despite the clear benefits of enrolling a significant percentage of employees fail to fill out the required paperwork.¹³⁴ One solution to this problem is automatic enrollment, whereby employers automatically enroll their employees in a retirement plan while making it easy for them to opt out. By exploiting the status quo (or default) bias, choice architects increase savings rates and they do so in a way that does not interfere with their employees' freedom of choice—again, it is easy for employees to opt out of the savings plan.¹³⁵ What should we think about such a policy, which clearly respects employees' freedom of choice and yet relies on a method of influence that fails to track reasons?

First, we should ask whether it would be possible to achieve the same desirable outcome—increased savings rates—via a process of influence that better tracks reasons. For example, employers might employ what T&S call a policy of “forced choosing.” This involves requiring employees actively to choose whether or not to enroll in a savings plan—they are neither automatically enrolled nor automatically left out.¹³⁶ T&S discuss studies showing that forced choosing results in higher enrollment rates than systems in which the default rule directs 100% of their income to their paychecks (that is, systems where employees have actively to switch from being not enrolled—the default—to being

¹³⁴ *Nudge*, p. 107

¹³⁵ *Ibid.*, pp. 106-109. T&S also discuss automatic enrollment in “Libertarian Paternalism Is Not an Oxymoron,” pp. 1159-1160 (and throughout the article)

¹³⁶ *Ibid.*, pp. 109-110

enrolled), but lower enrollment rates than those achieved by automatic enrollment.¹³⁷ To fill the gap in enrollment rates between forced choosing and automatic enrollment, employers might supplement forced choosing with rational persuasion. For instance, employers might explain why it is so important to save money for retirement. They might show employees how devoting a modest percentage of their monthly income to savings can with time (and compound interest!) make a dramatic difference in their standard of living during their retirement years. They might provide helpful charts showing how some given amount of money grows over time, given conservative projections. As T&S note, defined contribution plans (such as 401(k)) can be very attractive, as “contributions are tax deductible, accumulations are tax deferred, and in many plans the employer matches at least part of the contributions of the employee...[t]his match is virtually free money.”¹³⁸ If the benefits of enrollment are significant enough to justify nudging employees to enroll, why should not employers come out and openly cite these benefits when discussing retirement options with their employees?

I believe this last question should be the principal guiding one in evaluating the permissibility of nudge-like influences. In the absence of a convincing justification for refraining from providing influencees with the reasons the influencer believes support the behavior, it is hard to see why choice architects should resort to nudges. If a choice architect believes there are considerations that would justify her structuring a choice situation in a particular way in order to achieve some desired outcome, then she should believe she is justified in making those considerations apparent to the people whose behavior she wishes to influence. By making these considerations apparent to the

¹³⁷ Ibid., pp. 110

¹³⁸ Ibid., p. 107

influencee the influencer intends not only that the influencee does what she has good reason to do, but also that she does what she has good reason to do for the right reason.

There are cases, though, where the provision of reasons would be ineffective. It may be impossible to convince a person who is not responsive to reason that he ought to do X, and yet a nudge may succeed in getting him to do X. In other cases, it may not be feasible to make apparent the considerations that support the behavior at which the choice architect is aiming—e.g., a cafeteria manager may not find it easy rationally to persuade high school students to choose a spinach salad rather than pizza. Or, it may actually be counterproductive to rely on rational persuasion. In such cases, the provision of reasons may have unintended consequences that undermine the aim of providing them. For example, if a patient always defers to her doctor's judgment, irrespective of the situation, she may be more likely to act in light of the right reasons via a nudge-like influence than she would be if her doctor sought rationally to persuade her.¹³⁹

The central conclusion I wish to draw from the preceding discussion is that although nudges are not always all-things-considered morally impermissible would-be nudgers must provide moral justification for resorting to this form of influence. The precise form such justification will take will vary case by case, depending on the morally relevant features of the situation as well as on the plausibility of the would-be nudger's explanation for why straightforward rational persuasion would not be appropriate. A would-be nudger will have to say whether or not the target of his intervention is reasons-responsive; she may have to explain how the goodness of the outcome of her intervention outweighs the wrongfulness of deliberately aiming to leave her nudgee detached from the

¹³⁹ I am imagining a situation in which the nudge—e.g., the framing effect—allows the patient to consider her options on her own, something she would not do if she were reflexively to defer to her doctor's judgment.

considerations that ought to govern her behavior; she may have to explain how a nudge would better respect the nudgee's autonomy than would an attempt at rational persuasion. There will be situations where the justification for resorting to nudges is insufficiently strong. In these cases, it is impermissible to resort to nudges. In other situations, the justification will be very strong, such that nudging is clearly permissible or even obligatory.

Bibliography

- Baron, Marcia. "Manipulativeness," *Proceedings and Addresses of the American Philosophical Association*, Vol. 77, No. 2, (2003): 37-54.
- Beauchamp, Tom and James Childress. *Principles of Biomedical Ethics*, 6th edition. Oxford University Press, 2009.
- Blumenthal-Barby, Jennifer and H. Burroughs. "Seeking Better Health Care Outcomes: The Ethics of Using the Nudge," *The American Journal of Bioethics*, 12:2, (2012): 1-10.
- Bovens, Luc. "The Ethics of *Nudge*," in Till Grune-Yanoff and S.O. Hansson (eds.), *Preference Change: Approaches from Philosophy, Economics and Psychology*. Theory and decision library A (42), (Springer, 2009): 207-219.
- Buss, Sarah. "Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints," *Ethics* 115 (2) (2005): 195-235.
- Cave, Eric. "What's Wrong with Motive Manipulation?" *Ethical Theory and Moral Practice*, Vol. 10, No. 2, (2007): 129-144.
- Christman, John. "Autonomy and Personal History," *Canadian Journal of Philosophy*, Vol. 21, No. 1 (1991): 1-24.
- Fischer, John Martin and Mark Ravizza. *Responsibility and Control*. Cambridge University Press, 1998.
- Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility," *The Journal of Philosophy*, Vol. 66, No. 23 (1969): 829-839.
- . "Freedom of the Will and the Concept of a Person," *Journal of Philosophy*, Vol. 68, No. 1 (1971): 5-20.
- Frosch, Dominick, et al. "Authoritarian Physicians and Patients' Fear of Being Labeled 'Difficult' Among Key Obstacles to Shared Decision Making," *Health Affairs*, vol. 31, no. 5, (2012), accessed March 20, 2013, doi: 10.1377/hlthaff.2011.0576.
- Greenspan, Patricia. "The Problem with Manipulation," *American Philosophical Quarterly*, Vol. 40, No. 2 (2003): 37-54.

- Grice, Paul. "Logic and Conversation," *Studies in the Way of Words*, Harvard University Press, 1989.
- Halpern, Scott, et al. "Default options in advance directives influence how patients set goals for end-of-life care," *Health Affairs*, (2013) 32:2408-417, doi: 10.1377/hlthaff.2012.0895 (accessed March 20, 2013).
- Hausman, Daniel and Brynn Welch. "Debate: To Nudge or Not to Nudge," *The Journal of Political Philosophy*, Vol. 18, No. 1, (2010): 123-136.
- Hill, Thomas. "Autonomy and Benevolent Lies," *Journal of Value Inquiry*, 18:251 (1984): 251-267.
- Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias," *Journal of Economic Perspectives*, 5(1), (1991): 193-206.
- Kant, Immanuel. *Grounding for the Metaphysics of Morals*, trans. James W. Ellington, Hackett Publishing Company, 1993.
- Kligman, Michael and Culver, Charles. "An Analysis of Interpersonal Manipulation," *The Journal of Medicine and Philosophy*, 17, (1992): 173-197.
- Mandava, Amulya and Joseph Millum, "Manipulation in the Enrollment of Research Participants," *Hastings Center Report* 43, no. 2 (2013): 38-47. DOI: 10.1002/hast.144.
- Markovits, Julia. "Acting for the Right Reasons," *Philosophical Review*, Vol. 119, No. 2, (2010): 201-242.
- Mele, Alfred, *Autonomous Agents*, Oxford University Press, 1995.
- Mills, Claudia. "Politics and Manipulation," *Social Theory and Practice*, Vol 21, No. 1 (1995): 97-112.
- NIH Office of Behavioral and Social Science Research. Available at http://obssr.od.nih.gov/scientific_areas/health_behaviour/behaviour_changes/index.aspx (accessed March 20, 2013)
- Noggle, Robert, "Manipulative Actions: A Conceptual and Moral Analysis," *American Philosophical Quarterly*, Vol. 33, No. 1 (1996): 43-55.
- Nozick, Robert. "Coercion," in *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, Sidney Morgenbesser, Patrick Suppes, and Morton White (eds.), New York: St. Martin's Press, (1969): 440-472

- . *Anarchy, State, and Utopia*, Basic Books, 1974.
- Parfit, Derek and John Broome. "Reasons and Motivation," *Proceedings of the Aristotelian Society, Supplementary Volumes* Vol. 71, (1997): 99-146.
- Raz, Joseph. *The Morality of Freedom*, Oxford: Clarendon Press, 1986
- Redelmeier, D., P. Rozin, and D. Kahneman. "Understanding Patients' Decisions: Cognitive and Emotional Perspectives," *JAMA* 270, 1, (1993): 72-76.
- Saghai, Yashar, "Salvaging the Concept of Nudge," *Journal of Medical Ethics*, Published Online First: February 20, 2013, doi:10.1136/medethics-2012-100727
- Scanlon, Thomas. *What We Owe to Each Other*. The Belknap Press of Harvard University Press, 1998
- Stern, Lawrence. "Freedom, Blame, and Moral Community," *The Journal of Philosophy*, Vol. 71, No. 3, (1974): 72-84.
- Thaler, Richard and Cass Sunstein. "Libertarian Paternalism is Not an Oxymoron," *The University of Chicago Law Review*, Vol. 70, No. 4, (2003): 1159-1202
- . *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, 2008
- Rovane, Carol, *The Bounds of Agency*, Princeton University Press, 1998
- Rudinow, Joel. "Manipulation," *Ethics*, Vol. 88, No. 4, (1978): 338-347
- Twain, Mark. *The Adventures of Huckleberry Finn*, New York: Oxford University Press, 1996
- UK Institute for Government and the Cabinet Office, "MINDSPACE: Influencing behavior change through public policy," <http://www.instituteforgovernment.org.uk/our-work/better-policy-making/mindspace-behavioural-economics> (accessed March 20, 2013)
- Wilkinson, T.M.. "Nudging and Manipulation," *Political Studies*, article first published online September 7, 2012, doi: 10.1111/j.1467-9248.2012.00974.x
- Williams, Bernard. *Ethics and Limits of Philosophy*, Harvard University Press, 1986