Moral Responsibility, Authenticity, and Education

Ishtiyaque Haji and
Stefaan E. Cuypers

Moral Responsibility, Authenticity, and Education

Routledge International Studies in the Philosophy of Education

1. Education and Work in Great Britain, Germany and Italy
Edited by A. Jobert, C. Marry, L. Tanguy and H. Rainbird

2. Education, Autonomy and Democratic Citizenship
Philosophy in a Changing World
Edited by David Bridges

3. The Philosophy of Human Learning
Christopher Winch

4. Education, Knowledge and Truth
Beyond the Postmodern Impasse
Edited by David Carr

5. Virtue Ethics and Moral Education
Edited by David Carr and Jan Steutel

6. Durkheim and Modern Education
Edited by Geoffrey Walford and W. S. F. Pickering

7. The Aims of Education
Edited by Roger Marples

8. Education in Morality
J. Mark Halstead and Terence H. McLaughlin

9. Lyotard: Just Education
Edited by Pradeep A Dhillon and Paul Standish

10. Derrida & Education
Edited by Gert J J Biesta and Denise Egéa-Kuehne

11. Education, Work and Social Capital
Towards a New Conception of Vocational Education
Christopher Winch

12. Philosophical Discussion in Moral Education
The Community of Ethical Inquiry
Tim Sprod

13. Methods in the Philosophy of Education
Frieda Heyting, Dieter Lenzen and John White

14. Life, Work and Learning
Practice in Postmoderniity
David Beckett and Paul Hager

15. Education, Autonomy and Critical Thinking
Christopher Winch

16. Anarchism and Education
A Philosophical Perspective
Judith Suissa

17. Cultural Diversity, Liberal Pluralism and Schools
Isaiah Berlin and Education
Neil Burtonwood

18. Levinas and Education
At the Intersection of Faith and Reason
Edited by Denise Egéa-Kuehne

19. Moral Responsibility, Authenticity, and Education

Ishtiyaque Haji and Stefaan E. Cuypers

Moral Responsibility, Authenticity, and Education

Ishtiyaque Haji and
Stefaan E. Cuypers

Routledge
Taylor & Francis Group
New York London

First published 2008
by Routledge
270 Madison Ave, New York, NY 10016

Simultaneously published in the UK
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

This edition published in the Taylor & Francis e-Library, 2008.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk."

Library of Congress Cataloging-in-Publication Data

Haji, Ishtiyaque.
Moral responsibility, authenticity, and education / by Ishtiyaque Haji and Stefaan E. Cuypers.
p. cm. — (Routledge international studies in the philosophy of education ; 19)
Includes bibliographical references and index.
ISBN-13: 978-0-415-96468-5 (hbk)
ISBN-10: 0-415-96468-7 (hbk)
1. Responsibility. 2. Authenticity (Philosophy) 3. Education—Philosophy.
I. Cuypers, Stefaan E., 1958- II. Title.

BJ1451.H35 2008
370.11'4—dc22 2007050755

ISBN 0-203-89514-2 Master e-book ISBN

ISBN10: 0-415-96468-7 (hbk)
ISBN10: 0-203-89514-2 (ebk)

ISBN13: 978-0-415-96468-5 (hbk)
ISBN13: 978-0-203-89514-6 (ebk)

Ish dedicates this book to Shaheen
Stefaan to his Four Loves: Ann, Paulien, Marjan and Louise

Contents

Acknowledgments xi

1 Introduction: The Metaphysics of Responsibility and Philosophy of Education 1

2 Moral Responsibility, Authenticity, and the Problem of Manipulation 8

3 A Novel Perspective on the Problem of Authenticity 15

4 Forward-Looking Authenticity in the Internalism/ Externalism Debate 42

5 Authentic Education, Indoctrination, and Moral Responsibility 62

6 Moral Responsibility, Hard Incompatibilism, and Interpersonal Relationships 89

7 On the Significance of Moral Responsibility and Love 107

8 Love, Commendability, and Moral Obligation 127

9 Love, Determinism, and Normative Education 156

Appendices 189
Notes 223
References 233
Index 243

Acknowledgments

Work on this book commenced in the autumn of 2003 at the *Catholic University of Leuven,* Belgium. We are grateful to Arnold Burms for the insightful exchanges on a number of topics that we address in this book. We thank the Research Fund of the *Catholic University of Leuven* for its grants (F/02/072; F/05/077) which made possible our working together on this book. Ish Haji thanks the *Calgary Institute for the Humanities* for a fellowship that freed up time for him to work on the book.

We are especially indebted to Harry Frankfurt, Michael McKenna, Alfred Mele, and Derk Pereboom with whom we have discussed, on many, many, occasions, issues in the free will debate, and to David Carr, John Haldane, the late Terence McLaughlin, and Harvey Siegel for numerous conversations about topics in the philosophy of education that figure centrally in this work.

We would like, too, to express our gratitude to a number of philosophers who, over the years, have encouraged or assisted us in our work in the metaphysics of free will and moral responsibility, or in the philosophy of education, or in ethical theory: Randolph Clarke, Doret De Ruyter, Fred Feldman, John Martin Fischer, Robert Kane, James Marschall, Paul Smeyers, Ben Spiecker, Paul Standish, Jan Steutel, Wouter van Haaften, John White, Christopher Winch, David Zimmerman, Michael J. Zimmerman.

Parts of the following previously published articles of ours appear either verbatim or modified in the book. With permission of Routledge, Taylor & Francis Group: Ishtiyaque Haji & Stefaan E. Cuypers, "Moral Responsibility and the Problem of Manipulation Reconsidered," *International Journal of Philosophical Studies* 12 (2004), pp. 439–464; and Haji & Cuypers, "Magical Agents, Global Induction, and the Internalism/Externalism Debate," *Australasian Journal of Philosophy* 85 (2007), pp. 343–371; see, also, these Journals' web site: http://www.informaworld.com. With kind permission of Springer Science and Business Media: Haji & Cuypers, "Hard- and Soft-Line Responses to Pereboom's Four-Case Manipulation Argument," *Acta Analytica. International Periodical for Philosophy in the Analytical Tradition* 21:4 (2006), pp. 19–35; and Cuypers, "The Trouble with Externalist Compatibilist Autonomy," *Philosophical Studies. An International Journal*

for Philosophy in the Analytic Tradition 129 (2006), pp. 171–196. With permission of Blackwell Publishing: Haji & Cuypers, "Moral Responsibility, Love, and Authenticity," *Journal of Social Philosophy* 36 (2005), pp. 106–126; Cuypers & Haji, "Education for Critical Thinking: Can It Be Non-Indoctrinative?," *Educational Philosophy and Theory* 38 (2006), pp. 723–743; and Cuypers & Haji, "Authentic Education and Moral Responsibility," *Journal of Applied Philosophy* 24 (2007), pp. 78–94. With permission of the editorial board of the *European Journal for Analytic Philosophy:* Haji & Cuypers, "Love Imperiled," *European Journal for Analytic Philosophy* 3:1 (2007), pp. 5–20. We are grateful to the editors and publishers for their permission to use material from these essays.

Many thanks to Judy Ann Levine for proofreading the manuscript and her helpful recommendations, to Elizabeth Levine, Editorial Assistant at Routledge for her diligent work, and to Benjamin Holtzman, Research Editor at Routledge; their valuable contributions are very much appreciated.

1 Introduction
The Metaphysics of Responsibility and Philosophy of Education

1.1. THE ISSUES

This book addresses issues at the intersection of the metaphysics of freedom and moral responsibility, on the one hand, and the philosophy of education, on the other. Three related predicaments spur the inquiry. The first is a quandary in the free will debate. Moral responsibility requires a characteristic sort of freedom: add to a set of necessary and otherwise sufficient conditions for responsibility, the requisite freedom or control condition, and the set of conditions suffices for moral responsibility. Leading proposed accounts of such freedom run afoul of the notorious "manipulation problem." Various species of manipulation, such as unsolicited implantation of germane beliefs and desires in an agent, seem to undermine the agent's accessibility to moral praise or moral blame for conduct that causally stems from the implanted elements. Skinner's fictional character, Frazier, who is the founder of the imaginary, utopian world, Walden Two, crisply exposes what lies at the core of the nagging challenge of responsibility-subversive manipulation. Frazier explains that members of his community enjoy the freedom to do whatever they decide but all their goals, values, and desires have been conditioned into them in early childhood (Skinner 1948/1976). Whatever the control denizens of this utopian world exercise over their behavior, they are not morally responsible for this behavior because it is the causal output of desires, beliefs, and the like that are "alien" to them. Citizens of Walden Two, with their fabricated psychologies, are not the originators of their causal springs of action. In the sorts of case of interest, the aptly manipulated agent's choices are not free, again in the pertinent sense of "free" that moral responsibility presupposes, because these choices issue from elements that are inauthentic or not "the agent's own."[1]

If, however, we grant that manipulation of this sort undermines responsibility, how is such manipulation relevantly different from other causal forces having to do with, say, "normal" upbringing, that produce the same results as the manipulation but without allegedly undermining freedom?[2] Just as a resident of Walden Two has no control over the ultimate sources of his behavior—these would include his desires, values, and so forth that

have been implanted in him—so, it seems, persons with ordinary nurturing have no ultimate control over the acquisition of a significant subset of their motivational springs. How, then, can we ever be morally responsible for any of our behavior?

Determinism is, roughly, the view that all events, including everything that we do, are causal upshots of the distant past and the laws of nature. The nagging problem of manipulation has momentous import because it afflicts both accounts of the freedom responsibility requires which are compatible with determinism (compatibilist accounts) and those which are incompatible with determinism (libertarian accounts included).[3]

Next, consider a somewhat parallel problem in the philosophy of education. Proponents of otherwise diverse perceptions on the overarching goals of education concur that one of education's primary aims is to ensure that our children become *moral agents*.[4] In the pertinent sense of "moral agent," to be such an agent is, essentially, to be a competent partaker in the range of practices constitutive of moral responsibility. It is widely acknowledged that whereas certain forms of instruction or upbringing facilitate achieving this goal, various forms of paternalism or indoctrination impede or altogether derail its attainment. We submit that when such things as indoctrination jeopardize securing this goal, they do so because they imperil the child's progression into a moral agent; the victimized child may never evolve into an individual who is a suitable candidate for moral praise- and moral blameworthiness.

The second problem we tackle—"the problem of educational authenticity"—is now easily grasped. Necessarily, education involves "interferences" because it is a process of molding or shaping; it requires inculcating in the child, among other things, action-producing elements such as desires, deliberative principles, and values that will non-trivially influence the child's later reflections, choices, and overt conduct. But if such elements are implanted—the child's capacities of reflective control not in any way engaged in the acquisition of these elements because these capacities are nonexistent in the child at this early stage—is the child not relevantly like a puppet on a string akin to the inhabitants of Walden Two? Are these instilled elements not just as foreign to the child as are those with which the populace of Walden Two find themselves, or those a cult leader finagles into youngsters? The urgent concern to which sundry theorists of education call attention is that, as the requisite, pertinent educational interferences seem no different in kind than those of responsibility-subversive manipulation to which we have called attention, such interferences are incompatible with nurturing the child into a moral agent. Hence, an "authentic education," it is claimed, is a will-o'-the-wisp.[5]

As for the third problem, Harvey Siegel and others champion the view that an ideal of education is to ensure that our progeny develop into critical thinkers: children should grow into agents who can assess beliefs, desires, actions, reasons, and other pertinent psychological elements on the basis

of appropriate evaluative standards, be disposed to such evaluation, and be motivated by good reasons in belief-formation and action.[6] We are partisans both of this ideal and the related one that our children succeed in becoming *autonomous* critical thinkers.[7] This second ideal requires that the child mature into an agent who is self-governing with respect to the motivational constituents of being a critical thinker, such as the desire to assess reasons. The "indoctrination objection," however, calls into question whether education, aimed at cultivating autonomous critical thinkers, is possible. This time, the fundamental concern is that even nascent dispositions to evaluate reasons have not taken root in the young child. Thus, if the child is to turn into a critical thinker, the motivational building blocks of critical thinking must be "indoctrinated" into the child. Herein lies the rub: indoctrination flies in the face of autonomy, curtailing or altogether foiling development of the child into a *self-governing* agent.

Our initial, steering, objective is to propose a unified solution—a solution whose key components turn out to be more or less the same—to these problems. The solution unfolds in two stages. First, a person's candidacy for responsibility ascriptions calls for the person's being an agent of a suitable sort. We offer a partial analysis of this species of agency. Second, an individual is responsible for her behavior only if the behavior causally issues from motivational springs that are authentic. We defend an analysis of when such springs are authentic by focusing on features of causal pathways to the acquisition of these springs.

Regarding the former, a person must be a *morally normative agent* if she is to be morally responsible for her behavior. The effective capacity for full-fledged intentional deliberative action—the mark of a morally normative agent—requires possession of an evaluative scheme. Four constituents structure such a scheme: (a) Normative standards the agent believes should be utilized to assess reasons for action or to evaluate beliefs about how choices should be made. To be a fitting candidate for *moral* responsibility, the normative standards must include a set of moral principles or norms. (b) The long-term ends the agent regards as "direction-setting." (c) Deliberative principles the agent utilizes to arrive at practical judgments about what to do or how to act. (d) Motivation both to act on one's normative standards and to pursue one's long-term goals on the basis partly of one's deliberative principles. We propose that it is sufficient for an individual to be a morally normative agent at a time for that individual to have at that time an evaluative scheme with these four elements—the agent is minimally morally competent; has deliberative skills and capacities; and is able to act on the basis of at least some of her intentions, decisions, or choices.

Regarding authenticity, our view is that there is nothing like authenticity *per se;* motivational elements, such as desires, that are part of a person's evaluative scheme are not authentic in their own right. Rather, we defend a relational view of authenticity according to which motivational (and other) springs of action are authentic or inauthentic only relative to whether later

behavior that issues from these springs is behavior for which its agent is responsible. Elaborating, an answer to the three inaugural problems requires differentiating between causal routes to the acquisition of salient action-producing elements, such as desires and beliefs, which are *normal* and causal routes which are *deviant* relative to the ones that are normal. To isolate normal causal routes, in turn, we distinguish between two stages in an individual's life: the stage prior to which the individual has acquired an evaluative scheme—roughly, the phase of early childhood—and the stage after initial scheme acquisition. We argue for the view that the constituent elements of the child's initial evaluative scheme are relationally authentic in the manner just specified: they are authentic relative to ensuring (later) moral (or some other normative variety) of responsibility. A normal causal route to the acquisition of the elements of an initial evaluative scheme is subsequently specified in terms of the conditions for initial scheme authenticity. Constituents of one's evaluative scheme during the period following initial scheme acquisition are authentic insofar as they causally derive from modifications to the elements of one's initial scheme that one "freely" initiates; one undertakes the revisions under one's own steam.

The problem of educational authenticity is solved, in an analogous fashion, by invoking the view that authenticity *per se* of an initial scheme's constituents is a myth—we can succeed in turning children into morally normative agents only if appropriate desires, beliefs, values, and other things are "implanted," their implantation being crucial to the child's development into a responsible (normative) agent; and showing that things such as extreme paternalism and offensive indoctrination, unlike what are taken to be the "normal" sorts of thing that must be done to acquire salient action-producing elements, involve causal routes that are deviant when they are responsibility-undermining. Analogously, to dissolve the indoctrination objection, we suggest that the motivational prerequisites of being a critical thinker, such as the desire to acquire beliefs on the basis of pertinent evidence, even if instilled at a stage at which the child has insufficiently developed cognitive capacities, can be "truly the child's own" or autonomous only relationally: autonomous motivational elements are ones with respect to which the future child is self-governing.

As a bridge to an inquiry into pertinent issues of love, our second principal objective is to question the uncritically adopted assumption that education's chief, overarching goal is to secure the child's passage into a *morally* responsible agent. A seemingly compelling rationale for this assumption is that moral responsibility is vital largely because the costs to us of being without it are substantial. We challenge this rationale on the basis that the importance of moral responsibility in our lives has been overestimated. Other considerations, such as those of love, are just as or even more fundamental. This undertaking requires, among other things, clarifying the notion of importance in question, exploring why love is valuable, and responding to arguments that attempt to show that acting from moral duty,

which generally goes hand-in-hand with being morally responsible, is no different than acting from love, endeavors in which we engage in the latter half of the book.

Entertain the hypothesis, which we attempt to substantiate, that love is of paramount significance. An agent can act from love without morally "deontic" considerations—those of moral obligation, right, or wrong—playing any role in the generation of her love-dictated actions, such as behaving in a manner in which one takes on the concerns of the beloved even at significant costs to oneself. In such cases, the agent need not be *morally* praiseworthy for doing what love requires but may, nevertheless, be commendable from the standpoint of love, *commendability* being a species of *normative* and not merely causal responsibility that is not moral. In other cases in which, for example, the requirements of love and morality conflict, having discharged what she takes to be her moral obligation, the agent may not be morally blameworthy for her behavior but may still be censurable from the point of view of love, again, *censurability* being a variety of blameworthiness that is non-moral.

If it is appraisals of love, including those of commendability or censurability from love's standpoint, that are of primary importance in day-to-day living, then it would seem that one focus of education should be to ensure that our children turn into agents who are apt candidates for appraisals of love. Our relational account of authenticity is amenable to accommodating this recommendation. It stands to reason that children cannot unfold into agents who are fitting candidates for appraisals of love without various educational interferences. Such interferences, along the lines we have suggested, would be relationally authentic; other interferences that subvert later appraisability from the standpoint of love—commendability or censurability—would be relationally inauthentic.

Our discussion on love bears, among other things, on evaluating the intriguing proposal which, for example, Derk Pereboom has advanced, that living without free will is not as damaging as it has been made out to be.[8] A central strand of Pereboom's thought is that we value various aspects of love. Hard incompatibilism is, roughly, the view that, with the exception of agent-causal accounts of freedom, both compatibilist and libertarian accounts are incompatible with free action. According to Pereboom, hard incompatibilism leaves intact prominent aspects of love (and aspects of other reactive attitudes) that we value, so hard incompatibilism should not be so unsettling after all. This line of reasoning, however, is vulnerable to objection. First, if hard incompatibilism undermines the freedom of our decisions, it also undermines the freedom of affections such as emotional states. To the extent that interpersonal relationships are bound up with *free* emotional states, to that extent hard incompatibilism imperils them. Second, if hard incompatibilism undermines moral praise- and blameworthiness, it should equally undermine commendability and censurability. We argue that what we deeply value in lovable behavior is inextricably

associated with our being commendable for that behavior. Hence, hard incompatibilism undermines a deeply cherished (typical) constituent of loving relations—lovable behavior—*if* it undermines moral praise- and blameworthiness.

We conclude with two agendas for future inquiry. First, we have argued in a prior work that determinism undermines the truth of judgments to the effect that an action is morally right, wrong, or obligatory. This is because the truth of these "morally deontic" judgments presupposes that, when we act, we have the freedom to do otherwise but determinism, it seems, is incompatible with such freedom. We propose that this sort of argument, suitably adapted, may be toothless if invoked to impugn the requirements or prohibitions of love. Second, we take tentative steps to show how love contributes to the intrinsic value of a life for the person who lives the life. Our view is that (typically) we take delight in concerns of the heart; we take attitudinal pleasure in the fact that we act from love when we so act. Intrinsic attitudinal pleasures (and intrinsic displeasures) are prime contributors to the intrinsic value of lives.

In sum, the book progresses from an analysis of normative agency and authenticity, and a discussion of the relevance of these analyses to the manipulation problem and to pertinently related problems in the philosophy of education, to a defense of the thesis that responsibility from love's standpoint is of vital significance, and the implications of this thesis for what we deem to be legitimate goals of education and for other issues in the free will debate.

1.2. PROSPECTUS

The book is organized as follows. Chapter 2 expands on the manipulation quandary. Chapter 3 introduces our relational account of authenticity and applies it to this quandary. There are two appendices to Chapter 3. In one of these, we focus on other responses that have been proposed to the manipulation problem. We compare our response to these other responses. In the second appendix, we discuss an objection by Michael McKenna to the sort of response that we give—what McKenna dubs a "soft-line response"—to the manipulation quandry. Chapter 4 addresses the concern that the historical genesis of one's springs of action—how one acquires these springs—is largely irrelevant to whether one is morally responsible for actions that causally issue from these springs. Chapter 5 lays out the problems of educational authenticity and indoctrination in the philosophy of education, and shows that the solution to the quandary of manipulation expounded in the third chapter can be ably adapted as a solution to these problems as well. Chapter 6 explores the freedom of affective states and raises provisional doubts about the survival of sundry interpersonal relationships in a hard incompatibilist world. Chapter 7 argues for the relative

insignificance of moral responsibility and the relative importance of commendability and censurability (praise- and blameworthiness, respectively, from love's standpoint). A pivotal thesis introduced and defended in this chapter is the thesis that the value of loving behavior to us is essentially a function of our being commendable for the behavior. Chapter 8 defends this thesis against objections, and argues for the view that one may act from love without acting from duty and vice versa. Appealing to the thesis introduced in Chapter 7 and defended in the ensuing chapter, Chapter 9 reassesses the view that determinism leaves intact relations of love even if it undermines moral responsibility. The chapter also sketches how love may contribute to the intrinsic value of a life for a person, and, this in turn, suggests inroads into gauging the importance of what have been proposed as the various, overarching aims of education.

The core of our views is presented in the *chapters*. The *appendices* contain what we think are important, fairly closely related, but perhaps somewhat peripheral matters that we wish to discuss. Readers can choose to skip this material if it is irrelevant to their interests.

2 Moral Responsibility, Authenticity, and the Problem of Manipulation

2.1. INTRODUCTION: CONDITIONS OF RESPONSIBILITY

In the venerable tradition of responsibility that traces to Aristotle's *Nicomachean Ethics,* ignorance and lack of freedom can undermine a person's accessibility to moral blame or praise.[1] These widely accepted excusing conditions require supplementation with two others. One pertains to agency. We are exempt from responsibility if we fail to be agents of a certain sort. For example, if we are unable to regard any consideration as a reason for action, we cannot be morally blame- or praiseworthy for our behavior. Regarding some factor as a reason for an action requires an ability to see that, because of that factor, practical reason recommends performing the action. Or, again, if we are unable to evaluate reasons and judge, in light of our reasons, which course of action is subjectively best—best from the perspective of our own values—then we cannot be responsible. A second condition may be dubbed the "inauthenticity" condition. Its underlying idea is that one cannot be responsible for an action causally generated by actional springs such as desires, beliefs, or values that are not "truly one's own" or "inauthentic." An unwitting victim of brainwashing, having been "endowed" with a fresh set of values, goals, and other pro-attitudes, may willfully perform an anticipated transgression upon being released from captivity. Still, despite being an appropriate agent for responsibility ascriptions, having "responsibility-grounding" control in performing her action, and failing to act "out of" germane ignorance in doing what she does, many would agree that she is not deserving of blame for at least her *initial* offense; she is not blameworthy because she acted on actional springs that are not "authentic."[2]

Contemporary accounts of responsibility have striven to uncover and clarify the positive requirements of responsibility aligned with these four excusing conditions. The analysis we favor is that a person is morally responsible for performing an action if and only if he is an agent of an appropriate sort, he performs the action on the basis of the belief that he is doing something morally obligatory, right, or wrong, he has responsibility-grounding control in performing the action, and the action causally issues from authentic actional springs. Needless to say, each of these conditions

requires considerable elaboration and defense.[3] It is the last of these four conditions that is of immediate concern to us.[4] Our interest in the authenticity condition resides in the condition's intimate association with the formidable problem of manipulation.

2.2. CNC MANIPULATION AND THE AUTHENTICITY REQUIREMENT

Determinism is the thesis that, at any instant, there is exactly one physically possible future (van Inwagen 1983, p. 3). Compatibilism is the view that determinism is compatible with free action and moral responsibility. Incompatibilism is the denial of compatibilism. Unlike their traditional predecessors who hold that freedom and responsibility require alternative possibilities—these things require the freedom to do otherwise—conventional compatibilists by and large eschew this condition of control and argue, instead, for the replacement that free action must causally issue from appropriately structured psychological elements of a mentally healthy and competent agent. On some views, for instance, a free action derives from a first-order desire with which its agent identifies (Frankfurt 1971/1988); on others, a free action arises from a suitably reasons-sensitive process of deliberation (Wallace 1994; Fischer and Ravizza 1998; Haji 1998), where neither the hierarchical control nor the reasons-responsiveness at issue entails the freedom to do otherwise. Use the label "directional control" as a generic tag for the kind of control or freedom conventional compatibilists believe responsibility requires. However sophisticated their account of directional control, it has been touted that all varieties of compatibilism fall prey to the manipulation problem: evil neurologists or their likes may manipulate an agent, in the absence of the agent's awareness of being so manipulated, in such a fashion that the relevant psychological elements of the agent exemplify the structure required for free action. Intuitively, such agents are mere marionettes of their manipulators and hence, do not act freely or are not morally responsible for behavior that causally issues from their corrupted psychologies.

A particularly poignant incarnation of the manipulation (or "source"[5]) problem is the problem of "covert and nonconstraining (CNC) control." Introducing the problem, Robert Kane writes:

> In the case of constraining control, controlled agents are knowingly forced to do something against their wills. They are held at gunpoint or threatened with punishment if they do not do the controller's bidding, or they are locked in a room and simply prevented from doing what they want to do. . . . Nonconstraining (NC) control is another matter. It is exemplified by the cases of behavioral conditioning and behind-the-scenes manipulation. . . . In such cases, the controllers do

> not get their way by constraining or coercing others against their wills, but rather by manipulating the wills of others so that the others (willingly) do what the controllers desire. . . . In the most interesting cases, such control is a "covert" nonconstraining control—or CNC control, . . . in which the controlled agents are unaware of being manipulated or perhaps even unaware of the existence of their controllers. (Kane 1996, pp. 64–65)

Kane remarks that, "Frazier, the fictional founder of Skinner's Walden Two, gives a clear description of CNC control when he says that in his community persons can do whatever they want or choose, but they have been conditioned since childhood to want and choose only what they can have or do" (Kane 1996, p. 65). Assume that Wally, a member of Walden Two, is molded to be the sort of person that he is. His beliefs, values, goals, and so forth have been implanted in him. Since these elements that play an ineliminable role in his behavior are not "authentic"—Wally cannot claim "ownership" for them because they originate in sources beyond his control—he is not responsible for behavior that causally issues from these elements.

The possibility of being subjected to covert nonconstraining control is a key factor that impels Kane toward incompatibilism. Briefly, Kane argues that if we allow for an agent's being morally responsible for a causally determined action, we will also have to say that, despite being manipulated in ways that obviously subvert responsibility—despite, for example, being the victim of CNC control—agents who are so manipulated are, nevertheless, responsible. Kane's view is that the compatibilist cannot distinguish, in a principled fashion, between deterministic causal histories and responsibility-subverting manipulated causal histories.

The problem of CNC manipulation has, however, wider scope than has generally been acknowledged. If the problem detrimentally affects compatibilist accounts of responsibility or freedom, it seems equally to affect accounts of responsibility or freedom not consistent with determinism. Libertarians are incompatibilists who hold that at least some of us, at times, perform free actions for which we are responsible. Consider, for example, modest libertarians who adopt, more or less wholesale, a compatibilist account of directional control and then initiate a modification in the account by stipulating that, at some point or points along the causal pathway to the action for which an agent is responsible, the causal relation among elements that give rise to the action is nondeterministic (perhaps the causation is probabilistic). On one modest libertarian view, an agent's prior reasons to do something nondeterministically give rise to a decision to do that thing. Had the agent formed a decision to do something else instead, a different set of reasons would have nondeterministically given rise to that decision (Kane 1996; Ekstrom 2000). If a decision that deterministically arises from psychological elements that have been implanted in an agent is not one for which the agent is responsible, it is not clear why the agent should bear

responsibility for that decision *merely* in virtue of that decision's nondeterministically issuing from the implanted elements. The chance that the agent could have decided otherwise, given the nondeterministic causation of the decision, seems irrelevant to the concern: adroit enough manipulation could surely constrain the set of alternatives psychologically feasible for the agent. Or, consider, for example, a more robust type of libertarianism. Libertarians who are agent-causalists insist that an action—mental or otherwise—is free only if it is agent caused. Agent causation is a species of causation not reducible to ordinary garden-variety event causation. As its name connotes, if a person were to agent-cause a decision of hers, she as a substance, as contrasted to events involving her, such as her having of reasons, would be the first relatum of the causal relation that gives rise to her decision (O'Connor 2000; Clarke 2003). Agent causalists, just like compatibilists and event-causal libertarians who claim that the control free action and responsibility require consists, at least partly, in one's actions being nondeterministically caused by one's prior reason states, insist that reasons influence choice or intentional action that is free. Hence, if the agent's reasons have a pronounced effect upon her agent-causal activity, and these reasons are susceptible to manipulation, it should come as no surprise that the manipulation problem cannot be skirted merely by requiring that a choice or action be free only if it is agent caused. Libertarians, then, it appears, countenance a quandary of manipulation as pressing as the quandary compatibilists face.[6]

To motivate the view that at the heart of the manipulation problem is the condition of responsibility's requiring conduct that issues from authentic springs of action, ponder this case involving "global CNC manipulation." Imagine that neurology and neurosurgery have so progressed that not only can particular pro-attitudes such as desires, volitions, intentions, or goals be induced in an individual with or without the individual's consent or knowledge, but where one individual can be molded psychologically to be just the kind of person the surgeon desires. Jenny is a shy, unassuming woman with no family and friends. She lacks outstanding skills or distinguished capacities, and if she were to disappear from her workplace or domicile in Brooklyn, preliminary inquiries would be made only to be quickly suspended. Jamie, an accomplished computer hacker, has successfully masterminded several lucrative "hacking" offenses. Max, the shady entrepreneurial neurologist, eager to test a new form of psychosurgery, which if successful, will be used to "recruit" personnel, kidnaps and anesthetizes Jenny. During her sleep, Max works on Jenny, turning her, *in relevant respects*—specifically, in respects concerning her computing skills and work habits—into a psychological twin of Jamie. Flown to Brussels, Jenny is to begin work at Maxwell Incorporated, Max's computing firm. Max is the sole person aware of Jenny's transformation. Having settled Jenny into her new abode, he is killed on his return to New York. Knowledge of Jenny's transformation is buried with him.

Recovering from the surgery, Jenny has no suspicions that she has fallen victim to Max. She awakens with profound changes that, from her own inner perspective, she cannot but accept. The psychosurgery has endowed her with a new set of *pertinent* values, goals, preferences, and the like, while "erasing" ones discordant with these new elements that she formerly had. Assume that these implanted elements are practically unsheddable. As Alfred Mele (1995, p. 172) explains, an "actional" element such as a desire or belief is practically unsheddable for a person at a time if, given her psychological constitution at that time, ridding herself of that element is not a "psychologically genuine option" under any but extraordinary circumstances. Catching the morning news, Jenny learns about the new computing system installed in Barclays Bank, and after some diligent work, manages to transfer from an account in that bank a sizable sum of money into Maxwell's holdings. Although he will never know it, Max's transformation surgery has been a stellar success.[7]

This case of global CNC manipulation ("Psychohacker") assumes that pre- and post-surgery Jenny is identical. Understandably, some might be troubled about this assumption. A few refinements should pacify the skeptics. Suppose pre-surgery Jenny and Jamie do share certain types of goals, values, and preferences. Assume that the psychosurgery leaves intact in pre-surgery Jenny these shared elements. Suppose, in addition, that some of pre-surgery Jenny's goals, preferences, values, and other things which "compete" with those of Jamie's are also left intact together with some of pre-surgery Jenny's memories. Label the set of pre-surgery Jenny's competing psychological elements left intact the "minimum competing set." Assume, in addition, that the members of the minimum competing set, in conjunction with the shared elements, is the minimal cluster of psychological elements required to preserve personal identity so that we can be assured that pre-surgery Jenny *is* identical to post-surgery Jenny. Finally, assume that the memories of pre-surgery Jenny's that are left intact and all her competing psychological elements that are members of her minimum competing set are "repressed"; although post-surgery Jenny does have them, it is not possible, unlike pre-surgery Jenny, for her to "access" them. It would seem that under these conditions pre- and post-surgery Jenny are identical.[8]

The details of this case are consistent with assuming that Jenny exercises responsibility-grounding *control* in stealing, and that when she steals, she does not steal in relevant ignorance but on the basis of the belief that she is doing wrong in stealing. Still, it is intuitively implausible to regard Jenny as being to blame for her inauguratory criminal act. We suggest that she has an excuse because this action causally derives from engineered-in antecedents that are not "truly her own."

Sundry theorists about responsibility (or autonomy) concede and indeed defend this sort of claim. So, for instance, commenting on his own case of global manipulation in which Beth, a philosophy professor, is implanted with the psychological personality of Charles Manson, Mele writes:

> [E]ven though Beth is a psychological twin of Manson . . . , it does not follow that she autonomously possesses her Mansonian values. One indication of this is that, given the details of the case, we would not hold her *responsible* for her Mansonian character. Our reason for withholding attribution of responsibility (while supposing that Manson, her psychological twin, is responsible for his character) can only be that Beth was compelled to possess . . . her corrupt Mansonian values. . . . Manson, on our suppositions, is not relevantly different internally, but he autonomously possesses his values. (Mele 1995, p. 159)

Further, Mele declares that, on his view, manipulated Beth would not be responsible for the actions that flow from a character that was engineered in her against her will.[9] In addition, John Martin Fischer, Mark Ravizza, and Don Locke, for example, advance and defend analogous judgments about responsibility for actions that issue from an "implanted character."[10]

There are, of course, prominent accounts of responsibility that yield a contrary verdict. A first-order desire is, loosely, a desire whose object is some action. A second-order desire is a desire whose representational content is some actual or possible desire of its agent. Harry Frankfurt's hierarchical theory, very roughly holds that, assuming epistemic conditions of responsibility satisfied, a person is morally responsible for an action that issues from a first-order desire with which she identifies. A person identifies with a first-order desire if (on one version of the theory) there is an appropriate "fit" between an unopposed second-order volition of hers—an unopposed second-order desire regarding which first-order desire should move her to action—and the first-order desire that does in fact move her to action (Frankfurt 1971/1988). Relevant to our concerns is that this sort of hierarchical mesh condition *can* be satisfied by any number of acts of Jenny's, including her embezzlement, when she awakens from psychosurgery. The hierarchical theory, unlike Mele's views, would then yield the result that Jenny bears moral responsibility for her offense. There are, though, various problems with the hierarchical theory.[11] This is certainly not the appropriate place to document and assess them. Suffice it to say that even incompatibilists about determinism and responsibility, such as Kane, regard cases of global manipulation as posing a serious challenge to hierarchical approaches to responsibility (Kane 1996, pp. 66–67).

Still, the disagreement that some have with the moral we draw from cases such as Psychohacker, that there is a requirement of authenticity for responsibility (or autonomy), should not be lightly ignored. Opponents of the requirement will not regard post-surgery Jenny as having a legitimate excuse. They may argue that how an agent acquires her springs of action—whether the springs are implanted by covert psychosurgery or are "culturally nurtured"—need have no bearing on whether the agent is responsible for actions that appropriately derive from these springs. In addition to responding directly to an argument of this genre (in Chapter 4),

we suggest that one way to mediate this dispute is to develop an account of authenticity and expose its advantages. To the extent that the account is theoretically or explanatorily fecund, the more confident we shall be that the authenticity requirement is a bona fide requirement. We argue that, over and above generating intuitively satisfactory results in cases such as Psychohacker and helping to assess (and deflate) what appear to be compelling objections against compatibilist views of freedom, our account of authenticity enables us to resolve two distinct but related problems in the philosophy of education. The first is the problem of educational authenticity. Ensuring that our children develop into responsible agents, a fundamental goal of education, requires deliberate interferences in the processes that shape the child. However, these interferences seem no different in kind than interferences, such as Pavlovian conditioning or extreme paternalism, which appear to subvert agency. How, then, can instilled salient action-producing psychological elements of the child be "authentic" or, alternatively, how is an authentic education possible? The second is the problem of indoctrination which calls into question whether education, aimed at cultivating critical thinkers, is possible. The core of the concern is that since the young child lacks even nominal capacities for assessing reasons, the constituent components of critical thinking have to be indoctrinated if there is to be any hope of the child's attaining the ideal. In Siegel's words, if education for critical thinking is necessarily indoctrinative, "the ideal becomes significantly tarnished" (Siegel 1988, p. 78). We address these problems in the philosophy of education (in Chapter 5). First, though, we turn (in the next chapter) to developing our account of authenticity and to tracing its implications for cases involving manipulation.

3 A Novel Perspective on the Problem of Authenticity

3.1. INTRODUCTION: BASELINE RESTRUCTURING

The problem of manipulation can profitably be construed as a problem of deviance. In troubling cases of manipulation, psychological elements such as desires and beliefs, among other things, are acquired via causal routes that are deviant relative to causal routes deemed normal or, as we abridge, relative to causal routes that are *baseline*. Reconceptualizing the problem in this manner makes it more tractable. It challenges all parties to come to grips with what baseline is being presupposed either in a positive account of authentic springs of action—conative or doxastic elements that play an action-producing role and that arise in typical responsibility non-subverting fashion via "innocuous" causal routes—or in denunciations of either a proposed account of authenticity, or some compatibilist or incompatibilist characterization of control. In addition, lucidly articulating the baseline assumed, either in a criticism or constructive proposal, facilitates evaluation of that baseline.

Restructuring the manipulation problem as proposed has other advantages. On standard characterizations of the problem, one may inadvertently give the impression that a concern about manipulation arises only in instances of purposeful or intentional mischief by third parties.[1] Plainly, though, this need not be so. Inheriting a gene whose phenotypic expression, let us assume, ensures that an agent cannot discern right from wrong, undermines freedom or responsibility. Appealing to baselines, we are not barred from supposing that certain causal routes to acquiring such things as desires that do not involve purposeful determination are, at least intuitively, deviant.

In addition, it is fairly customary to delineate inauthentic springs of action by citing a list of factors, such as coercion, hypnosis, and indoctrination, that, if part of the etiology of the springs or otherwise properly associated with them, undermine their authenticity. Our account that invokes baselines identifies a common feature underlying these heterogeneous factors: causal routes to acquiring actional springs incorporating such factors (if they are indeed responsibility-undermining) are not "normal" or baseline.

In what follows, we endeavor to make headway in meeting the manipulation problem. We distinguish baseline causal routes to the acquisition of

germane conative or cognitive action-producing psychological elements from deviant ones. We begin with a suggestion regarding a salient feature of baselines that is incompatibilist—roughly, the view that a causal route to the acquisition of germane actional elements is baseline only if the agent has ultimate control over the acquirement of these elements. Although we believe that this incompatibilist candidate fails, it has the virtue of bringing into relief two different strategies to isolate and defend baselines, one of which is enlisted in this work. We then propose and defend rudiments of what we deem is a baseline acceptable independently of whether one has compatibilist or incompatibilist leanings.

3.2. AN INCOMPATIBILIST CANDIDATE

Kane, a libertarian, has advanced a powerful attack against compatibilist accounts of freedom. As previously noted, he explains that an agent who is covertly and nonconstrainingly ("CNC") controlled is unaware of being so controlled, and the controllers get their way, not by constraining or coercing the agent against her will, but by manipulating her so that she willingly does what the controllers desire (Kane 1996, p. 65). The controllers may achieve their end, for example, by judiciously and surreptitiously implanting suitable desires and beliefs. Kane argues that a CNC-controlled agent is not free; her ends or purposes not being her own, she lacks the control over her behavior that responsibility requires. Such an agent, he proposes, on any compatibilist account of freedom, may well be free. A deterministic acquisition of conative and doxastic elements compatibilists regard as "normal" is in no relevant manner any different than acquisition of such elements by covert and nonconstraining manipulation; nature plays the role of the evil demon or neurologist. Hence, Kane concludes, if covert and nonconstraining manipulation undermines responsibility-relevant freedom—if the acquisition of pertinent springs of action as a result of such manipulation is deviant—so is the acquisition of these springs as a result of the deterministic unfolding of history.

We will revert to the interesting premise that compatibilist accounts imply that CNC-controlled agents lack the freedom responsibility requires. For our immediate concerns, we elaborate the prior premise to uncover an incompatibilist condition for baselines. Kane suggests that the CNC-controlled agent is not ultimately responsible for her behavior because she lacks ultimate control over that behavior. Indeed, he defines *free will* as the power to be the ultimate creators (or originators) and sustainers of our own ends and purposes (Kane 1996, p. 4). In a world in which God pre-sets all of the reasons, motives, and purposes of agents, the causal route to the acquisition of these things is deviant *because* the agents are not ultimately responsible for them. "Their wills," Kane claims, "in the form of their reasons, motives, and purposes are already 'set one way' before they act

and they are never the ultimate sources of their own wills" and subsequent behavior (Kane 2000*a,* p. 69). Kane's remarks suggest the following condition on baselines.

> *Baseline UC:* A causal route to the acquisition of salient action-producing elements, such as desires, beliefs, best judgments, intentions and so on, is baseline ("normal") only if the agent has ultimate control over their acquirement.

The condition requires disambiguation as there are at least two different conceptions of ultimate control, one negative, the other positive. Ultimate control is concerned with forging an intimate link between an agent's putatively free action and the agent herself so that it is, minimally, plausible to maintain that the agent is the "final" source of her action. Assume that any free action is caused.[2] The two conceptions of ultimate control that are relevant to our discussion share the following: (i) The cause, or at least a causal antecedent, of the free action must be a component of the type of cause that plays a salient role in the production of action or free action (such as the having of a suitable belief or desire). (ii) This cause (or part of it) must, in some obvious sense, be internal to its agent. (iii) The cause must be at least partly constitutive of the agent in a way in which, in virtue of being so constitutive, it would be correct to say that the action (or the free action) "truly" issues *from the agent* or is the "*agent's own*." One type of compatibilist, for instance, who claims that free actions causally arise from first-order desires with which we identify—first-order desires appropriately associated with higher-order psychological elements of ours—may accept these three conditions as sufficient for ultimate control.[3] However, no libertarian would do so unless the causal relatum of the action that meets these three conditions satisfies some further condition. A libertarian who endorses the *negative* conception of ultimate control conceives of this cause as an event (or state) and adds to the trio of conditions that this cause *not* be causally determined if it deterministically gives rise to the action or free action, or it nondeterministically produces the action or free action. Kane, for example, contends that ultimate responsibility for an action requires either that the action not be causally determined by its causal antecedents, or if the action is so causally determined, any determining cause of it be the result of some other action of the agent that was not causally determined and for which the agent is ultimately responsible (Kane 1996, p. 35).[4]

The *positive* conception of ultimate control adds to (i), (ii), and (iii) the additional condition that the action (or free action) be agent caused. Recall, agent-causal accounts of free action typically maintain that the freedom moral responsibility requires is to be explained in terms of agents possessing causal powers to perform actions or to make choices without being causally determined to do so. On these views, the variety of causation free action requires is not reducible to causation among ordinary events. Rather,

the sort of causation is an instance of a substance—the agent—directly causing a choice or a causal precursor of the choice but not by way of any states or events.

Common to all forms of agent-causal accounts of free action are two themes. First, an agent who agent causes her action exerts greater causal control in its production than do deterministic and libertarian counterparts who are not agent causes. This is because agents who agent cause their free behavior, it is suggested, exercise what can be dubbed "dual-directional control" over their putatively free actions. Let us say that an agent has *genuine alternatives* at a time just in case, consistent with the past and the natural laws remaining fixed, he could, at that time, have performed any of two or more alternatives. (Intentionally refraining from acting qualifies as an alternative.) An agent has *dual-directional control* if and only if, just prior to the moment at which he performs some action, he had genuine alternatives, and he "determines" which of these he performs—in some way, he directly controls the choice he makes. He has some further power to influence causally which of his alternatives he realizes, a power over and above the mere chance of acting differently.[5] Second, proponents of agent-causal accounts of free action claim that when an agent agent causes a free action, she herself is an uncaused cause of that action. In this way, she is the ultimate source, and, hence, an ultimate originator of her action.

Reconsider *Baseline UC,* the suggestion that a causal route to the acquisition of salient action-producing elements is baseline only if the agent has ultimate control over the acquirement of these elements. If it is negative ultimate control to which this condition appeals, the condition falls prey to Kane's own worries regarding the adequacy of compatibilist accounts of freedom. As we previously explained, the mere absence of deterministic causation or the introduction of nondeterministic causation among relevant action-producing conative and doxastic elements, provides modest libertarianism with no iron-clad security against concerns of covert and nonconstraining manipulation.[6]

The positive conception of ultimate control, in contrast, is more promising, at least initially. Suppose, then, that baselines require that agents acquire salient action-producing elements either by direct agent-causal activity, for instance, agent causing an intention to do something, or by indirect agent-causal activity, for example, forming a belief as a result of prior deliberation that itself involves direct agent-causal activity. This incompatibilist contender suggests two different approaches to developing baselines. The first involves isolating a certain capacity whose exercise by the agent results in "authentic" salient action-producing elements. With the particular proposal under consideration, the capacity is the capacity of the agent to agent cause her behavior. The second approach involves isolating some factor that *cannot* be causally induced.[7] Again, with the proposal at issue, it has been argued that the complex event, *an agent's agent causing a (particular) action,* cannot itself be caused.[8] The following criticism of the

incompatibilist contender, to which we formerly eluded and now develop, calls into question the theoretical value of the second approach.

Revert to Psychohacker which we introduced in the previous chapter, the case of global manipulation in which Max turns Jenny into a psychological twin of Jamie. We said that it seems reasonable to judge that Jenny is not morally responsible for at least the first few actions after her surgery (actions that causally derive from her engineered-in values), a verdict we doubt would be affected if it were supposed that prior reasons of hers nondeterministically caused these actions. Then, it seems, stipulating that Jenny agent causes her initial decisions should make no difference either to this verdict: she is still not morally responsible for them. Typically, agent causalists embrace the view that reasons influence intentional action that is free. Timothy O'Connor, for example, claims that recognizing "a reason to act induces or elevates an objective propensity of the agent to initiate the behavior. . . . [M]y reasons structure my activity . . . in the more fine-grained manner of giving me, qua active [i.e., agent] cause, relative tendencies to act" (O'Connor 2000, p. 97). If the agent's reasons have this sort of effect upon her agent-causal activity, and manipulators can temper with these reasons, the manipulation problem cannot be evaded solely by requiring that an action is free only if it is agent caused. The fact, then, if it is one, that the complex event that is the event of an agent's agent causing an action has no cause provides no immunity against the manipulation problem.[9] We, thus, conclude that the second approach to developing baselines—isolating something that resists causal induction—is problematic. Our preference is to exploit the first approach that pays particular attention to certain capacities of the agent.

An action expresses a pro-attitude only if that pro-attitude plays a nondeviant causal role in the production of that action. We indicate that the judgment, if well-founded, that Jenny is not responsible for her actions that express her unsheddable, implanted actional springs in a version of the case in which her actions are deterministically caused, nondeterministically caused, or agent caused, presupposes some baseline relative to which the causal route via which her engineered-in values, desires, and other things are acquired is deviant.

Well, what *is* the normal causal route to the acquisition of salient action-producing elements relative to which other routes are deviant? In the rest of this chapter, we offer and defend a compatibilist contender.

3.3. A COMPATIBILIST CONTENDER

Some spade work will be helpful. Responsibility has freedom, epistemic, authenticity, and agency requirements. Here, we set aside the first two requirements and concentrate, instead, on the other two.[10] Beginning with the agency requirement, to be morally responsible, one must be an agent of

a certain sort, a "morally normative agent" as we have said. One agency requirement for responsibility is that the candidate be capable of intentional deliberative action. Such action, in turn, requires some psychological basis for evaluative reasoning. An agent's deliberations that issue in a practical judgment about what to do, which in turn gives rise to a decision or intention, involve the assessment of reasons for or against action by appeal to the agent's evaluative scheme. Such a scheme is made up of the following constituents: (a) Normative standards the agent believes (though not necessarily consciously so) ought to be invoked in assessing reasons for action, or in evaluating beliefs about how the agent should go about making choices. To be an apt candidate for *moral* responsibility, the normative standards must include a set of moral principles or norms; the agent must be minimally morally competent. She must understand the concepts of rightness, obligatoriness, or wrongness, and she must be able to appraise, morally, various choices or actions in light of the moral norms that are elements of her evaluative scheme. There is no requirement that appraisals be fully considered, free of error, or even conscious. Nor is there any requirement that the norms are evidentially based or justified in any strong sense of "justification." The agent may simply assimilate, without critical scrutiny, various norms of her religion or culture. (b) The agent's long-term ends or goals he deems worthwhile or valuable. Arnold, for example, may underscore his commitment to attempt to maximize overall happiness whenever he acts. (c) Deliberative principles the agent utilizes to arrive at practical judgments about what to do or how to act. For instance, Arnold may believe that the best way to maximize utility is to rely on rules of thumb like "keep your promises," "don't cheat," "don't steal," and other things. (d) Lastly, motivation both to act on the normative standards specified in (a) and to pursue one's goals of the sort described in (b) at least partly on the basis of engaging the deliberative principles outlined in (c).

We propose that it is a *sufficient* condition of an individual's being a morally normative agent—an appropriate candidate of moral responsibility—at a time, *t,* if that individual has at *t:* (i) an evaluative scheme with the requisite moral elements—the agent is minimally morally competent; (ii) deliberative skills and capacities; for example, the agent has the capacity to apply the normative standards that are elements of its evaluative scheme to evaluating reasons; and (iii) executive capacities—the agent is able to act on at least *some* of its intentions, decisions, or choices. An individual, like a toddler, who fails to have deliberative or executive capacities, will be able to exert much less control, if any, over its actions than an individual who does have such capacities. Read condition (ii) to entail that the agent is (at *t*) able to engage in genuine deliberation; her deliberative activities must meet the threshold of rationality below which such activities fail to count as bona fide deliberation.

Next, to develop the authenticity requirement and defend a compatibilist baseline, we begin with distinguishing two different stages in an individual's

life. The distinction coincides, roughly, with the margin between childhood and adulthood: differentiate the stage before the individual has acquired an evaluative scheme from the stage after which the individual's initial scheme has been acquired.[11] As standardly interpreted, global manipulation cases, such as Psychohacker or Mele's Beth/Manson mind experiment, involve manipulation at a stage after the individual has acquired an initial scheme. To understand when manipulation subverts moral responsibility at this stage, it will be helpful to understand when it subverts responsibility at the first stage.

We outline our overall picture of evaluative scheme authenticity. We then explain the connection between this picture and causal routes to the acquisition of salient action-producing elements that are baseline. Regarding authenticity of evaluative schemes of "developing agents" like us, we start with the following preliminary idea and then refine it. As a child matures, the child acquires an evaluative scheme; the child becomes a normative agent. The child's *initial evaluative scheme* is the scheme that the child initially acquires. If the constituents of such a scheme are properly acquired, in a sense of "proper acquisition" to be explained, then these constituents are authentic, and the initial scheme is authentic. Over time, the initial scheme evolves; its constituents change. Suppose we can give an outline of when, for instance, the desires and beliefs that are parts of an individual's initial scheme are authentic if they are authentic. We may then hypothesize that the desires and beliefs of an evolved scheme are authentic provided that they causally derive from modifications to the individual's initial scheme that are acceptable, again in a sense of "acceptable" to be supplied. A scheme that results from acceptable modifications to an individual's authentic initial scheme is that individual's *authentic evolved scheme*. Our guiding idea is that an agent's evaluative scheme is authentic if it is either the agent's authentic initial scheme or it is the agent's authentic evolved scheme.

Evaluative schemes contain both doxastic and motivational elements. Addressing the states before and after an evaluative scheme has been attained, is there a reasonable sense in which an agent's motivational and doxastic elements constitutive of the scheme, either initial or evolved, that the agent will acquire are authentic? In this book, we confine discussion primarily to motivational constituents. It is profitable to distinguish between *what* action-producing motivational element is instilled and its *mode* of instilment. An inept manipulator may implant a desire that the agent can easily thwart with the result that the manipulation is benign. An implanted desire that is irresistible, in contrast, may well subvert responsibility. Regarding mode of acquisition, "acquisition" is used broadly so that a desire instilled by third-parties counts as an acquired pro-attitude. However, so does a pro-attitude with which the child finds herself as a result of her genetic endowment.

Young children are not normative agents.[12] Because of this, they are not morally responsible for their behavior. Is there still, though, a sense

in which some of their beliefs and desires are authentic whereas others are not at the pre-normative agent stage? Reflect on mental illness or coercion, factors frequently thought to affect responsibility. Such factors subvert responsibility, when they do, if they undermine one or more of the requirements of responsibility such as epistemic or freedom requirements.[13] With this as our cue, we propose that *a pro-attitude or its mode of acquisition is inauthentic if that pro-attitude or the way in which it is acquired subverts moral responsibility for behavior, which owes its proximal causal genesis to the pro-attitude* (typically in conjunction with other springs of action), *of the normative agent into whom the child develops*. Subversion of moral responsibility would occur as a result of either epistemic, control, or other necessary requirements (*independent,* of course, of agency presuppositions) of moral responsibility being thwarted. In this sense, there is nothing like "plain authenticity" or "authenticity *per se,*" but only "authenticity with an eye toward responsibility," or "responsibility-relative authenticity." Preliminary comments on this view should, we anticipate, be illuminating.

First, some may propose that "plain authenticity" is a real possibility. It may be suggested that a desire is plainly authentic if it is smoothly integrated with the other pro-attitudes, or more generally, psychological elements of the agent. Or a desire consistent with, or partly constitutive of, one's acting in a manner in which one is "true to oneself" is plainly authentic. The suggestion, though, falls to serious doubt.[14] Caught at a vulnerable age, a young member of a cult, due to the cult leader's insidious indoctrination, may acquire desires that complement the victim's psychology and that define her character. However given their causal history, we would be reluctant to regard these desires as authentic, however integrated they might be with the other elements of the agent's psychological make-up.

Second, one might, again, wonder about the precise connection between authenticity and moral responsibility. A powerful intuition cases involving CNC control, such as Psychohacker, elicit is that the agent is *not* responsible partly but pivotally because she acts on springs of action "not truly her own," or, in our terminology, "not authentic." Our concern is to give a partial account of *this* sense of "authenticity."

The connection between authenticity and responsibility can be highlighted in a different fashion. One of the primary aims of educating children is to ensure that they become moral agents, and we have said that what it is to be a moral agent, in the germane sense of 'moral agent,' is to be a competent participant in the family of practices constitutive of moral responsibility. Among other things, to become a moral agent, the child must see herself as an appropriate candidate of the reactive attitudes such as indignation, resentment, and love and must be such a candidate.[15] It is accepted wisdom that whereas certain forms of training or upbringing are conducive to attaining this goal, various forms of paternalism or indoctrination are detrimental to its realization. We suggest that paternalism or indoctrination threaten attainment of this goal, when they do, primarily

in virtue of the fact that they threaten achievement of the desideratum that the child will be an apt candidate of things like praise and blame. Indoctrination and paternalism of the relevant sort thwart this fundamental goal of moral education *because* the severely afflicted child may not be a *moral person*—a morally normative agent—in contradistinction to a mere human being; indoctrination or paternalism foil the complex, intentional process—the bread and butter of moral education—of transforming a child from being simply a member of the species *homo sapiens* into a moral agent. We may, indeed, regard such indoctrination as *mere training* and not moral *education*.

Third, as many compatibilists and incompatibilists acknowledge, it is intuitively plausible to theorize that there is a connection between freedom and authenticity: behavior causally arising from springs of action not truly one's own is not behavior that is free in the sense of 'free' responsibility requires. We have proposed that plain authenticity is a myth; authenticity is relational. This relational conception is to be articulated in terms of a normal way of acquiring appropriate psychological elements. It follows that divergence from the normal pathway is related to lack of freedom. Should one find the view that there is a conceptual connection between freedom and normality suspect, consider this analogy: There is a conceptual connection between freedom and control; lack of control compromises responsibility-relevant freedom. Assume that the best account of control is an account that draws partially upon the notion of behavior's non-deviantly arising from causal springs not acquired as a result of one's normal mechanisms of deliberative control having been bypassed. It would follow that there is a conceptual connection between freedom and relevant bypassing, something not intuitively obvious.

To add more flesh to the relational account, some of Joel Feinberg's observations on child development are telling. Feinberg remarks that the extent of a child's role in his own shaping is a process of continuous growth begun at birth. He continues:

> From the very beginning that process is given its own distinctive slant by the influences of heredity and early environment. At a time so early that the questions of how to socialize and educate the child have not even arisen yet, the twig will be bent in a certain definite direction. . . . From the very beginning, then, the child must—inevitably *will*—have some input in his own shaping, the extent of which will grow continuously even as the child's character itself does. After that, the child can contribute towards the making of his own self and circumstances in ever increasing degree. These contributions are significant even though the child is in large part (especially in the earliest years) the product of external influences over which he has no control, and his original motivational structure is something he just finds himself with, not something he consciously creates. Always the self that contributes to the

> making of the newer self is the product both of outside influences and an earlier self that was not quite as fully formed. That earlier self, in turn, was the product both of outside influences and a still earlier self that was still less fully formed and fixed, and so on, all the way back to infancy. At every subsequent stage the immature child plays a greater role in the creation of his own life, until at the arbitrarily fixed point of full maturity, he is at last fully in charge of himself. . . . Perhaps we are all self-made in the way just described, except those who have been severely manipulated, indoctrinated, or coerced throughout childhood. But the self we have created in this way for ourselves will not be an authentic self unless the habit of critical self-revision was implanted in us early by parents, educators, or peers, and strengthened by our own constant exercise of it (Feinberg 1986, p. 34–35).

In this insightful passage, Feinberg astutely suggests that authenticity requires both a certain sort of maturation—one free of things like indoctrination or coercion—*and* deliberate interferences in the processes that shape the child. He proposes, for instance, that the habit of critical self-revision must be *implanted* in us early if we are to acquire autonomy. On Feinberg's view, then, some deliberate interferences in shaping the child are perfectly compatible with and are, indeed, required for authenticity.[16]

Keeping in mind this view of Feinberg and the proposal that instilment of pro-attitudinal (and doxastic) elements that subvert responsibility for subsequent relevant behavior undermines authenticity of such elements, ponder these examples. We said that to be *morally* responsible for an action, an agent must have elementary moral concepts, such as those of wrong or obligation, and she must be able to appraise morally (even if imperfectly), decisions, actions, consequences of action, and other things in light of the moral norms that are partly constitutive of her evaluative scheme. With agents like us—human beings—a minimally morally competent agent has a grasp of the notions of guilt, resentment, praise-, and blameworthiness or of notions of related reactive attitudes or feelings, and has at least a rudimentary appreciation of when such attitudes or feelings are appropriate. Suppose a child, Youngster, is trained in such a fashion that she simply lacks knowledge of the relevant moral concepts so that she is not even minimally morally competent. Then failing to instill the appropriate moral concepts is responsibility subversive since the lack of them precludes her from having the epistemic capacity required for moral responsibility at later stages of her development. Or consider instilling in Youngster a pro-attitude or disposition, the influence of which on her behavior she simply cannot thwart. Instilling such a pro-attitude—an irresistible desire, for example—would presumably undermine responsibility for later conduct arising from that pro-attitude by undermining the control responsibility requires. Or suppose Youngster is instilled with a powerful disposition always to act impulsively. Here, again, we would not want to hold Youngster responsible for much of

her later impulsive behavior. Or, finally, consider an interference that prevents Youngster from engaging in critical self-reflection. This may subvert responsibility for some of Youngster's later behavior by significantly narrowing, on occasions of choice, the range of Youngster's options, a range she could, in all likelihood, have considered had she acquired "normal" habits of critical self-reflection.

Some interferences, then, where *interference* is a general term for things like suppression of innate propensities, or implantation of certain dispositions, or deliberate lack of instilment of various pro-attitudes, are incompatible with the agent's being appraisable for his subsequent behavior which issues from instilled elements; such interferences subvert later responsibility while others do not. We propose that the subversive ones are (morally) *responsibility-wise inauthentic*. Specifically, imagine an agent, like a young child, who does not yet have an initial scheme. Such an agent's, *S*, having pro-attitude *P* is responsibility-wise inauthentic if *S*'s having *P,* as a result of instilment, subverts *S* being morally responsible for *S*'s behavior that stems from *P;* having *P* precludes *S* from being morally responsible for behavior that stems from *P.* Setting aside agency requirements of responsibility, subversive interferences undermine later moral responsibility by undermining *other* requirements of responsibility, such as epistemic or freedom requirements.

We have, so far, limited discussion to responsibility-relevant authenticity of the "objects" of instilment such as dispositions or pro-attitudes in general. What about the modes of instilling such things; are some responsibility-wise authentic and others not? We can approach this issue in the following manner. Assume, to ensure prevention of subverting moral responsibility for later behavior, it is necessary to instill in the child the disposition to be moral. Different modes of instilling this disposition could affect responsibility-relevant authenticity of this very disposition itself. Suppose, for example, that given the mode of instilling the moral disposition in Youngster—perhaps the disposition was "beaten into" Youngster, or instilled via "shock therapy"—Youngster subsequently finds that she cannot refrain from doing what she perceives to be morally right. On occasions of choice, she is stricken with inward terror even at the faintest thought of not doing what she deems moral. Intuitively, Youngster would not be responsible for much of her later behavior because the mode of instilling the moral disposition subverts responsibility-grounding control. Modes of instilling pro-attitudes (habits, dispositions, etc.) are responsibility-wise not "truly one's own" (that is, are responsibility-wise inauthentic) if they subvert responsibility for later behavior. Again, putting aside agency demands of responsibility, if these modes of acquiring pro-attitudes undermine later moral responsibility, they will do so by subverting one or more of responsibility's requirements.

In addition to pro-attitudes, a person's evaluative scheme comprises *cognitive* constituents—beliefs about both normative standards for evaluating reasons for action and deliberative principles regarded as appropriate for

arriving at practical judgments about what to do or how to act in particular circumstances. With the young child whose evaluative scheme is in embryo, it may well be that certain beliefs will have to be willfully instilled to ensure responsibility-relative authenticity. Perhaps, as Feinberg suggests, one will have to instill in the child *the belief that critical self-evaluation is important;* without this belief, moral responsibility for later behavior may well be threatened in the manner previously indicated. Further, the child's having of such a belief, it would seem, would be morally permissible and perhaps even morally required. Instilling beliefs of this sort, in consequence, via modes that do not subvert later responsibility, would not threaten responsibility-relative authenticity. Various sorts of belief, though, *would* undermine or seriously imperil moral responsibility for later conduct. The following sorts, for example, seem to fit the bill: beliefs formed as a result of deception (and self-deception), beliefs formed on the basis of coercive persuasion, and deliberately implanted beliefs formed on the basis of processes that bypass ordinary mechanisms of belief formation—such as subtle conditioning or subliminal influencing—in cases in which the agent did not consent to the implantation. The agent, presumably, would not be morally responsible for actions performed in the light of such beliefs.

To help in formulating a general principle about initial scheme responsibility-relative authenticity, we introduce some terminology. We have suggested that, possibly, having some pro-attitudes (dispositions, etc.) is required to ensure moral responsibility for later behavior—having them ensures that necessary conditions other than agency ones of moral responsibility can (later) be satisfied by the agent or by her behavior that stems from them. Such required pro-attitudes and beliefs are "authenticity demanding." We have also suggested that the having of some pro-attitudes (dispositions, etc.) and beliefs is incompatible with moral responsibility for later behavior which issues from them; having them precludes satisfaction of necessary conditions other than agency ones, such as epistemic or control conditions, moral responsibility requires. Such incompatible pro-attitudes and beliefs are inauthentic and may be dubbed "authenticity destructive." Lastly, we have suggested that some modes of instilling pro-attitudes (dispositions, etc.) and beliefs are incompatible with moral responsibility for later behavior; such modes of instilment subvert later responsibility by thwarting satisfaction of necessary conditions of responsibility apart from agency ones. These irreconcilable methods are "authenticity subversive." Finally, we propose this principle as one that governs responsibility-relative authenticity of initial schemes of "developing agents":

> *Authenticity-1:* An agent's initial evaluative scheme is responsibility-wise authentic if its pro-attitudinal and cognitive constituents (i) include all those, if any, that are authenticity demanding; (ii) do not include any that are authenticity destructive; and (iii) have been acquired by methods not authenticity subversive.

In other words, *Authenticity-1* says the following. An agent's initial evaluative scheme is responsibility-wise authentic if its pro-attitudinal and cognitive elements (i) include all those, if any, that are required to ensure that the agent will be morally responsible for its future behavior; (ii) do not include any that will subvert the agent's being responsible for future behavior that issues from these elements; and (iii) have been acquired by means that, again, will not subvert the agent's being responsible for its future behavior. The crux of *Authenticity-1* reduces to this: the agent's, such as the child's, initial evaluative scheme is not the agent's own if its pro-attitudinal or cognitive elements subvert, to a substantial degree, moral responsibility for later behavior that issues from these elements. We provide this gloss of the sense of "issues from." A causal theory of action (which we endorse) assumes that actions causally (and non-deviantly) arise from desires, or desire/belief pairs, or a cluster of psychological elements. On this theory, when an action issues from a certain desire (as opposed to another), this desire (as opposed to the other, typically together with other actional elements) is causally implicated in the production of the action. We presuppose whatever account of "issues from" that causal theories of action presuppose.

We amplify this account by responding to the pronouncement that *Authenticity-1* is either *empty* or *circular*. Regarding the former, some may think that *Authenticity-1* is empty because we have left open what *all* the other requirements or dimensions of responsibility are, and of the dimensions that we have acknowledged—the epistemic and control dimensions—we have not said in what these consist. So the implication of *Authenticity-1* that if (during the pre-initial scheme stage), a pro-attitude, such as a desire, or its mode of acquisition, undermines future responsibility for behavior that issues from it, that pro-attitude is inauthentic, is empty.

In reply, setting aside responsibility's agency presuppositions, no one doubts that responsibility has additional requirements. There is (general) concurrence, for example, that there are epistemic and freedom demands on responsibility. That responsibility has various requirements cannot be denied even in the face of disagreement concerning precisely *what* these requirements might be. Are there, for instance, requirements *other than* agency, epistemic, and control requirements? Concerning a putative particular requirement, such as the control one, theorists may well disagree on the *substantive account* of this requirement. Yet again, though, free will theorists (generally) do *not* deny that there is *some correct rendition* of this requirement (whatever it may turn out to be). Reconsider the claim that a pro-attitude acquired during the pre-initial scheme stage is inauthentic if behavior that issues from it is behavior for which the agent is not morally responsible by virtue of this pro-attitude's undercutting one or more of responsibility's requirements. On the assumption that there is a *correct* account of *what* the requirements of responsibility are (with the exception, again, of agency requirements), and that there is a *correct*

account of what each of these requirements *consist in,* this claim is decidedly *not* empty.

Regarding the concern of circularity, the consideration that there is a fact of the matter about what the requirements of responsibility are, and that there is a fact of the matter about what each of responsibility's requirements amounts to, also vindicates the view that *Authenticity-1* is not circular. We do not explicate the notion of initial scheme authenticity by covert appeal to the authenticity of the scheme's constituents. Rather, we first emphasize that on our view, during the pre-initial scheme stage, there is nothing like authenticity *per se* of pro-attitudes acquired during this stage. The authenticity of pro-attitudes is forward-looking or relational, specified by way of a relation between the having of a pro-attitude at a time when the child is not yet a normative agent, and behavior that issues from this pro-attitude at a time when the child has turned into a normative agent. Second, we account, specifically, for the authenticity in question in terms of whether an agent's behavior that owes its proximal causal genesis to this pro-attitude (again, typically, in consort with other actional elements) is behavior for which the agent is morally responsible. We ask whether the pro-attitude subverts responsibility for such behavior by subverting one or more of the requirements (other than agency ones and *ipso facto* other than authenticity ones) of responsibility. Assuming that there *are* such requirements (whatever they may turn out to be), and that there is a *correct* account of what these requirements consist in, we see *no* circularity in *Authenticity-1*.

Authenticity-1, in turn, motivates the following baseline for times prior to which the child has acquired an initial scheme.[17]

> *Baseline-1:* A causal route to the acquisition of a pro-attitude, or more generally, salient action-producing elements, is baseline (normal) if these elements have been acquired by means that are not authenticity subversive, and either the elements are authenticity demanding or they are not authenticity destructive. If some salient action-producing element is not acquired via a baseline route, the route to its acquisition is deviant.

We are confident that the condition of not being acquired by means that are authenticity subversive and the condition of not being authenticity destructive are also necessary conditions of a causal route's being baseline (at the pre-initial scheme stage). Hence, if, for instance, certain pro-attitudes are beaten into Youngster, and for this reason these pro-attitudes subvert later responsibility for action that issues from them, the causal route to their acquisition is deviant.

Summarizing, roughly, a causal route to acquiring things like desires or beliefs is *Baseline-1* if their acquisition does not subvert responsibility for later behavior that (at least partly) issues from these elements by subverting epistemic or control requirements of responsibility. It is sufficient that

the causal route to the acquisition of a desire be deviant if, for instance, behavior that stems from the desire is behavior for which the agent is not responsible because the desire is irresistible.

Once an initial scheme is in place, the condition for a causal route's being baseline is governed primarily by the principle that deviant routes are causal pathways that result in salient action-producing elements being acquired independently of the agent's "engaging" her authentic initial scheme. Before the condition and the principle can be more fully articulated, we need to say something about when changes in an initial scheme preserve authenticity.

Assume that Youngster has acquired an authentic initial evaluative scheme. Evaluative schemes are not static but dynamic; they can evolve. So, for instance, Youngster can renounce values formerly cherished and acquire new ones; she might come to question her belief that moral decisions should conform to the teachings of her religion and adopt a utilitarian outlook; or she might give up her deliberative principle that she should review her decisions frequently before implementing them because she finds that frequent review in certain contexts hinders success. Some modifications or changes in one's evaluative scheme may be perfectly compatible with preserving responsibility-relative authenticity whereas others will subvert authenticity. To distinguish between the two sorts of change, we require a conception of *acceptable* modifications.

As one's evaluative scheme is comprised of doxastic and motivational constituents, changes in one's scheme can involve changes in one or both. The general rule for acceptable modifications in either of these types of constituents is straightforward: the modifications must be made under one's own deliberative control. With respect to changes in pro-attitudes such as desires, instilled ones or newly acquired ones are acceptable as long as the actions, if any, to which they give rise are ones over which the agent has responsibility-grounding control and the changes are initiated by the agent's "exercising" (or engaging) her initial scheme. The changes occur as a result of exploiting capacities, such as deliberative ones, that the agent (substantially) has in virtue of elements constitutive of her authentic scheme. We allow for cases in which, through a series of past steps over which one has responsibility-grounding control, one deliberately instills in one a pro-attitude (or a cluster of such attitudes) which will give rise to actions over which one will lack such control. For instance, a person desperate to quit smoking may have implanted in her an irresistible desire to avoid cigarettes. A global change in pro-attitudes, as in Jenny's case in Psychohacker, that destroys initial morally normative agency, or completely "represses" it, subverts authenticity; acquisition of the implanted pro-attitudes bypasses completely Jenny's capacities of deliberative control.

To elaborate, Mele (1995, pp. 166–72, 183–84) plausibly proposes that most normal, healthy human agents have the following capacities in some measure: the capacity to modify the strengths of their desires in the service

of their normative judgments, of aligning their emotions with relevant judgments, of mastering motivation that threatens to produce or sustain (biased) beliefs that would violate their principles for belief acquisition and belief retention, of rationally assessing their values and principles, of identifying with their values and principles on the basis of informed critical reflection, and of modifying their values and principles should they judge that this is called for. We do not see how it is possible to possess these capacities without having an evaluative scheme. One cannot assess values and principles, for instance, without embracing a set of normative principles. These principles are causally and, ideally, non-deviantly engaged in, for instance, one's appraisal of a thought that strikes one upon witnessing a disastrous event. If one cares deeply about another, one must be able to modulate, appropriately, favorable emotions in response to one's belief that the cared for has fared or will fare well, and a range of unfavorable emotions in response to one's belief that the cared for has fared or will fare poorly. One cannot do so without engaging elements of one's evaluative scheme. Here, again, there will presumably be an appropriate causal story to be told about how one brings one's relevant emotions (the having of which themselves depends upon constituents of one's evaluative scheme) in line with one's pertinent judgments. Consonant with what Mele says, we propose that an agent's evaluative scheme is not engaged in, for instance, acquiring a pro-attitude, if the acquisition of this pro-attitude bypasses all of the agent's capacities of deliberative control (see, for e.g., Mele 2006, pp. 166–67). The modification to one's evaluative scheme resulting from its supplementation with this pro-attitude is not an acceptable one. If, in acquiring a pro-attitude, one manifests deliberative control and, thus, in this sense "engages one's evaluative scheme," the degree of deliberative control that one exercises will be a function of a number of factors, such as whether the deliberative process involves certain sorts of inefficiency and irrationality, like various sorts of selective biasing, and the coming to mind, while deliberating, of irrelevant considerations or akratic influences.

Assuming an appropriate account of authenticity for doxastic elements, we now advance the following sufficient condition of when evaluative schemes of "developmental agents" like us are relationally authentic.

> *Authenticity-2:* If agent *S*'s evaluative scheme at a time, *t,* is either *S*'s initial responsibility-wise authentic scheme at that time, or is an evolved responsibility-wise authentic scheme of *S*'s at that time—it is a scheme resulting from *acceptable* modifications to a scheme possessed by *S* prior to *t* that is responsibility-wise authentic—then *S*'s evaluative scheme is responsibility-wise authentic.

Consider *Jimma's victimization*. Clusters of desires and beliefs are implanted in (adult) Jimma, without Jimma knowing anything about the implantation, that cause her to perform action *B*. The unsheddable implanted

desires are not irresistible, and when she does *B,* she has the right sort of control for moral responsibility in doing *B*. Nor does the manipulation undermine epistemic requirements of responsibility. The manipulation is proficient enough to ensure that even if Jimma reflects on her implanted actional elements, the reflection that sanctions the action recommended by these elements—action *B*—stems from *further* engineered-in actional elements. *B* is an action Jimma would not have performed but for the manipulation. Intuitively, she is not morally responsible for *B*-ing. This is because *B* causally issues from elements that are *not* part of Jimma's authentic evaluative scheme.

Suppose *E-Scheme* is Jimma's initial or evolved authentic evaluative scheme at time *t,* and that an action of Jimma's, *A,* stems from actional elements that are not part of that scheme at *t*. The notion of part-hood is perplexingly complex but the underlying guiding idea is clear-cut. The actional elements are not part of that scheme as they have not been acquired under Jimma's own steam. They have not, for example, been attained as a result of Jimma's practical deliberation on the basis of *E-Scheme*. Nor has Jimma given consent of any sort to having these elements implanted in her. Rather, the elements have been acquired independently of Jimma's "engaging" *E-Scheme*. Our suggestion for a baseline for times after which an agent has acquired an authentic evaluative scheme, whether initial or evolved, is the following.

> *Baseline-2:* A causal route to the acquisition of a pro-attitude, or more generally, salient action-producing elements, is baseline (normal) at *t* if these elements are acquired as a result of the agent's engaging or exercising the pertinent constituents of the agent's authentic evaluative scheme at *t*. If, for instance, a pro-attitude is not acquired via a baseline route, the route to its acquisition is deviant.

Baseline-2 underscores the requirement that there must be an appropriate causal connection between actional elements that are acquired at a time and the agent's authentic evaluative scheme at that time for the causal route to the acquisition of these elements to be normal or baseline. Although it is extremely challenging to spell out precisely the nuances of this connection, the general direction of our position should be fairly evident. In Jimma's case, for example, there is no appropriate causal connection between her authentic evaluative scheme and her implanted elements. Regarding these elements, her authentic evaluative scheme is on the sidelines and not, in any fashion, "engaged." Such disengagement suffices for the relevant causal route's being deviant. What qualifies as an appropriate causal connection will, partially, depend on biological, psychological, and neurophysiological discoveries. Presumably, a proximal intention produced by a neuroscientist's directly stimulating an agent's brain and one produced by the normal engagement of elements of the agent's evaluative scheme will have different

neurological pathways or signatures. Such differences will be significant to distinguishing baseline from deviant causal routes.[18]

3.4. OBJECTIONS AND REPLIES

We now defend our "baseline proposal" against various objections. We deflect an initial concern. It may be argued that externalists about the mental should insist that for *S* to possess the belief that it is permissible for *S* to steal, *S*'s acquisition of this belief must have the right causal genesis. The right causal history, in turn, includes factors in the external world and, thus, its elements are not limited to features internal to *S*'s psychology. Hence, it may be contended that externalists should reject the assumption that beliefs or psychological states can simply be implanted in a subject's mind. It is though, wishful thinking that the manipulation problem would simply evaporate if externalism regarding the mental were true. Surely, *S* could acquire the belief that it is permissible for *S* to steal as a result of indoctrination on the part of *S*'s cult leader. Presumably, beliefs acquired in this fashion are not authentic.

Another preliminary concern focuses on temporal factors regarding *Baseline-1*. It may be objected that since *Baseline-1* is forward-looking, it cannot be "applied" at the time when a child acquires an initial scheme. It is, thus, vacuous. Whether a causal route to the acquisition of a pro-attitude *P* at time *t* is baseline (normal) or deviant *can only be fathomed in the future,* at a time $t + n$ future to *t,* the assessment turning on whether the having of *P* subverts moral responsibility for behavior that issues from *P* at or subsequent to $t + n$. Hence, whether a causal route at *t* is normal or deviant cannot be established at *t* itself since the correct verdict awaits what the relevant facts concerning responsibility for behavior stemming from *P* will be at $t + n$. How can the normality or deviancy of a causal route *now* depend on something that happens *later?*

In reply, we distinguish between two concerns about *Baseline-1*. The first is an epistemological one. Suppose desire *D* is acquired at *t*. How do we know at *t* that *D* is authentic at *t?* Unless we have foreknowledge, we do not know. However, this is not our concern. The second, which *is* our concern, is a metaphysical one. Suppose it is true that actions that later issue from *D* (at $t + n$) are actions for which we are not responsible. Then it is also true at *t* that actions which issue from *D* are ones for which we are not responsible. Hence, it is true at *t* that *D* is inauthentic. Because it is true at *t,* it is timelessly true—it is true at any time—that, at $t + n$, *D* gives rise to actions for which we are not responsible. Again, suppose that, at *t,* Pauline acquires the irresistible desire, *D*,* to do *A* at $t + n$. Pauline has no idea that *D** is irresistible at the time of its acquisition (at *t*). Yet *D** is responsibility-subverting at the time it is acquired. In short, the epistemological difficulty does not undermine *Baseline-1*.

Some might wonder whether there is a relevant empirically discernible difference at *t* between a responsibility-undermining desire *D1* and a normal (non-deviant) desire *D2,* each acquired at *t*. Whether there is such a difference between these two desires at the time of their acquisition is neither here nor there for our purposes. At *t, D1* (the inauthentic desire) has the property of *being such that actions that issue from it are ones for which the agent will not be responsible*. At *t, D2* lacks this property. This is the substantial difference between *D1* and D2 to which we call attention. *D1* need not "feel" any different than *D2* to its agent at *t*. Empirical considerations of this sort, as far as we can see, have no bearing on authenticity.

Kane, as we observed, suggests that any compatibilist account of freedom entails that a CNC-controlled agent may well be free and responsible. Similarly, McKenna wonders: why "couldn't a manipulator manipulate in a manner that facilitated the relevant authenticity-friendly agential abilities," thus undermining the agent's status as a morally responsible agent? (McKenna 2004*b,* supplement, sec. C) To assess this sort of concern, we distinguish a number of cases. In the first ("Manipulation 1"), assume that manipulation that involves, for example, instilling a set of desires, occurs at a time prior to the time at which an initial scheme is in place. The relevant question is whether the causal route to the acquisition of these desires is *Baseline-1-normal* or pertinently deviant. The route would be relevantly deviant if the implanted desires, for example, were to subvert responsibility for behavior that causally issues from them by undermining control or epistemic requirements of responsibility. If the manipulation leaves unscathed these requirements and, hence, fails to threaten the responsibility-relative authenticity of the implanted desires, then though such manipulation is perhaps morally improper, it should not subvert responsibility for behavior that issues from these desires. So manipulation via baseline causal routes, toothless as it is, is compatible with responsibility. Forms of manipulation such as CNC control that involve deviant causal routes are responsibility undermining. Just as there is no authenticity *per se,* there is no effectual manipulation *per se;* manipulation (in the context) is of concern only if it is manipulation relative to undermining responsibility for later behavior. The matter, though, is complicated because a person may be responsible for something to a certain degree; manipulation may curtail degree of responsibility without undermining responsibility altogether. Further, the acquisition of initial schemes, as Feinberg remarks, is not an all or nothing affair. A scheme may be partially acquired and the acquisition of its elements may involve numerous, diverse processes such as operant, aversive, and classical conditioning, role model imitation, and uncritical acceptance of the teachings of one's parents or the teachings of other authoritative figures. The more complete the initial scheme, the greater the extent to which the causal routes to implanted elements is deviant if the routes fail to "engage" parts of the scheme already in place.

Consider a second case ("Manipulation 2") in which manipulation "facilitates the acquisition of authenticity-friendly agential abilities." Imagine a world in which the pertinent agents, otherwise much like ordinary human beings, produce viable offspring only if various constituents of initial evaluative schemes are induced in the offspring. Given conditions at the world and the biological nature of its agents, if a segment of the evaluative scheme of newborns were not engineered in place shortly after their birth by established means, the newborns would not survive. The *minimal set* of actional springs is the set of pro-attitudinal and doxastic elements that require implantation to ensure survival of the progeny. In this scenario, we propose that for an initial scheme to be authentic, elements of the minimal set should respect *Baseline-1;* they should not subvert responsibility for later behavior these elements generate by undermining responsibility's epistemic or control requirements. In addition, if a pro-attitude is not a member of the minimal set and, if implanted, the pro-attitude would subvert responsibility for subsequent behavior that causally issues from it by subverting responsibility's control or epistemic requirements, it is inauthentic; the causal pathway to its acquisition is deviant. In this case, manipulation in a manner that facilitated authenticity-friendly agential abilities—the implantation of the minimal set—would not undermine responsibility.

Consider, finally, a third case ("Manipulation 3") in which, again, manipulation facilitates the acquisition of authenticity-friendly agential abilities. Reproduction in the imagined world—the *Blade Runner* world—requires that bodies, including brains, be fabricated. Once these things are prepared, close to *complete* initial schemes must be instilled if the "offspring" are to survive. Just as the ordinarily acquired initial schemes of humans can vary widely, so too can the instilled schemes of these humanoids. Somewhat akin to what we suggested in Manipulation 2, in this third case we propose the following. Baseline routes entail instilling pro-attitudinal and doxastic elements that mirror such elements of an ordinarily acquired authentic initial scheme in this respect: none of the instilled elements undermine responsibility for subsequent actions of the agent that derive from the instilled elements as a result of these elements subverting epistemic or control requirements of responsibility.

Suppose Alph is such a newly formed humanoid in the *Blade Runner* world. One might worry that Alph has been manipulated in a manner that has facilitated the relevant authenticity-friendly agential abilities; others have implanted most of his initial scheme and, hence, Alph is not responsible for the actions that issue from the implanted elements. We disagree. We do not know what baseline this concern *presupposes*. The baseline assumed cannot plausibly be the same as the one proposed for ordinary human beings. As our tale is spun, ordinary human beings enjoy a degree of latitude in the acquisition of initial schemes far greater than the degree our hypothetical humanoids enjoy in the acquisition of their initial schemes. Given the type of species, members of that species must be allowed as much

leeway as is possible, consistent with their biological natures, in the acquisition of initial schemes if instilling salient action-producing elements is not to undermine responsibility. If Alph enjoys the requisite latitude in acquiring his initial scheme, then nothing stands in the way of Alph's being an apt candidate of responsibility ascriptions.

The last two cases highlight an important moral. What is deviant is relative to what is normal and normality is normative. If the normal course of development of Alph-like creatures is a course of development as the one specified, then Alph's "manipulation"—the instilment of his initial scheme—is not responsibility subversive.

It will be instructive to revert briefly to the Walden Two case. Recall, Frazier, the founder of *Walden Two,* advances a suggestive case of responsibility-subversive manipulation when he says that in his community persons can do whatever they want or choose, but they have been conditioned in a way concealed from them since childhood to want and choose only what they can have or do. We initially remarked that whatever sort of control people in Walden Two exercise over conduct, it appears that they are not morally responsible for their behavior, again, because it is the causal output of desires, beliefs, values, and the like that are, in some manner, foreign to them. This case, however, depending on how its details are filled in, masks a complication. In one respect, the case may be likened to the case of children: to see that children develop into morally responsible agents, we "implant" various pro-attitudes into them. If the case is construed along these lines, then (as our relational view of authenticity implies) members of the utopian world may well *be* responsible for some of their conduct. Understood in a different fashion, though, their springs of action turn out to be foreign in, roughly, the way in which the springs of action of globally manipulated agents are foreign.

Various remarks of Tomis Kapitan suggest another set of objections. Kapitan proposes that a condition of adequacy of an account of free action—the *independence condition*—is that inasmuch as the requirements of responsibility, such as epistemic and control requirements, "provide non-question-begging criteria for judging whether [agent] S is responsible, then they must be decidable independently of assessments of S's worthiness of being praised or blamed" (Kapitan 2000, p. 83). Kapitan worries, first that, contrary to intuition, "the totally manipulated agent's scheme can be normative-wise authentic inasmuch as it is the agent's *initial* evaluative scheme" (Kapitan 2000, p. 101, n. 10). Depending upon the details of the manipulation, either this charge is innocuous or the intuition can be resisted. We indicate, again, that whether the intuition is persuasive depends upon an assumed or implicit baseline, something this first worry leaves unspecified. In addition, we have proposed that, with developmental agents like us, there is no such thing as initial scheme authenticity *per se;* how could there be any such thing? There is only responsibility-relative authenticity. Reflection on Alph's case suggests that, as long as the instilled

elements constitutive of Alph's initial scheme are not incompatible with responsibility for later behavior that issues from these elements, the initial scheme is not inauthentic; the causal route to the acquisition of these elements is (*Baseline-1*) normal.

Kapitan, worries, second, that reliance on the notion of responsibility in principles such as *Authenticity-1, Baseline-1,* and *Authenticity-2* threatens to violate the requirement that whatever the proposed conditions of responsibility, they "be decidable independently of assessments of responsibility" (Kapitan 2000, p. 101, n. 10). Kapitan claims that unless we know what the conditions of responsibility are, we cannot say whether the totally manipulated agent is responsibility-wise authentic (Kapitan 2000, p. 101, n. 10). We interpret Kapitan as proposing that without knowing what the conditions of responsibility are, we cannot determine whether a globally manipulated agent's evaluative scheme is authentic. The totally manipulated agent, he says, might satisfy the various conditions of responsibility, though intuitively, we may think it obvious that he is not responsible. However conditions such as *Authenticity-1* and the others "are not sufficiently independent to justify this claim" (Kapitan 2000, p. 101, n. 10).

In response, we do not see why principles such as *Authenticity-1, Authenticity-2, Baseline-1,* and *Baseline-2* are not relevantly independent. Again, in assessing various cases, we stress the importance of distinguishing the stage prior to which an agent has acquired an initial evaluative scheme and the stage after an initial scheme has been acquired. Contemplate for instance, a situation in which an irresistible desire is implanted in a child. Suppose actions that issue from this pro-attitude (along with other actional elements) are ones over which the child fails to have responsibility-grounding control (whatever the correct account of control). *Baseline-1* yields the result that the causal route to the acquisition of this desire is deviant. There is no reason why, in principle, similar assessments cannot be made in connection with "totally manipulated agents." To ascertain whether the requirements of *Baseline-1* are met in these cases, we appeal to *necessary* conditions of responsibility, to wit, control and epistemic conditions, and *not* sufficient conditions. Whether the control and epistemic conditions in these cases are satisfied may be determined independently of responsibility assessments: first, we decide whether these two individually necessary conditions are satisfied quite apart from assessing whether the agent is responsible; only *thereafter* do we judge whether the agent is also *responsible,* contingent upon whether the pertinent authenticity condition—*Authenticity-1*—is satisfied, in addition to the satisfaction of the two other conditions. Control, lack of ignorance, and authenticity are individually necessary but, at best, only jointly sufficient for responsibility. Hence, we fail to see why it is false that the desideratum that the requirements of *Baseline-1* "be decidable independently of assessments of . . . [the agent's] worthiness of being praised or blamed" is satisfied in the germane cases. *Baseline-1* seems not to violate the independence condition.

Reconsidering Jimma's predicament, owing to her pertinent action issuing from causal springs that are not part of her authentic scheme, she is not responsible for this action. *Baseline-2* gives the result that the causal route to Jimma's acquiring these motivational springs is deviant and, thus, that the causal route to this action from its motivational springs is deviant. Perhaps all of Jimma's actions have this sort of deviant causal history. Again, we do not appreciate why there is a concern with the independence condition in connection with *Baseline-2* even if Jimma were totally manipulated (at the post-initial scheme stage); whether the requirements of *Baseline-2* have been met can be ascertained by focusing on whether necessary conditions of responsibility are imperiled. When assessing *Baseline-2* requirements, we do not have to know *beforehand* whether the jointly sufficient conditions for responsibility are met: to judge whether the totally manipulated agent's evolved scheme satisfies *Authenticity-2,* we only put to work the individually necessary control and epistemic conditions (or other necessary conditions, apart from the agency ones, should there be any).

Yet another objection is that there is no principled way to distinguish between cases of CNC control and those of mere causal determination; the two types of case are analogous with respect to all factors pertinent to responsibility. Hence, if CNC-controlled agents are not responsible for their behavior, causally determined agents should not be responsible either, no matter what compatibilist-friendly condition of control or other germane conditions of responsibility are advanced.[19] The primary problem with this objection is its failure to specify baselines both at the pre- and post-initial scheme stage relative to which causal routes to the acquisition of salient action-producing elements are deviant. Suppose a desire is instilled in Youngster at the pre-initial scheme stage by means that do not subvert responsibility for later behavior that derives from the desire. The causal route, whether deterministic or nondeterministic, to the acquisition of this desire is not deviant, given *Baseline-1*.

It may be rejoined that the desire was not acquired under Youngster's own steam but as a result of factors beyond her control. So how could the desire be authentic? There is though, little reason to be concerned with this complaint unless one presupposes that there is something like authenticity *per se* at the pre-initial scheme stage and, moreover, that the relevant baseline the complaint presupposes is defensible.

To vary the case, suppose that the acquisition of Youngster's desire is incompatible with responsibility for later behavior that issues from it (and, perhaps, other causal antecedents). On our proposal, the causal pathway to this desire's acquisition is deviant regardless of whether the pathway is deterministic or not. Similarly, at the post-initial scheme stage, *Baseline-2* gives us the wherewithal to distinguish between causal routes to the acquisition of salient action-producing elements that are tainted with covert and nonconstraining manipulation and those that involve mere causal determination. In Jimma's case, for instance, the action for which

Jimma is intuitively not responsible issues from actional elements Jimma bears, the causal routes to the acquisition of these elements, whether the causation involved in these routes is deterministic or nondeterministic, being deviant.

3.5. PEREBOOM'S FOUR-CASE ARGUMENT

In a different context of assessing compatibilism, Pereboom advanced a particularly puissant version of the last objection—whether it is possible to draw a principled distinction between cases that involve CNC (or relevantly similar) manipulation and cases that involve mere causal determination—that merits separate treatment. Pereboom proposes that any compatibilist account of free action and moral responsibility succumbs to a pressing problem, significant aspects of which involve manipulation. Presupposing determinism, Pereboom conjures a sequence of four cases in each of which Plum kills White for personal gain, and then asks where to draw the line between Plum's not being morally responsible and his being morally responsible for the murder. Here is the first case:

> **Case 1.** Professor Plum was created by neuroscientists, who can manipulate him directly through the use of radio-like technology, but he is as much like an ordinary human being as is possible, given this history. Suppose these neuroscientists "locally" manipulate him to undertake the process of reasoning by which his desires are brought about and modified—directly producing his every state from moment to moment. The neuroscientists manipulate him by, among other things, pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic. Plum is not constrained to act in the sense that he does not act because of an irresistible desire—the neuroscientists do not provide him with an irresistible desire—and he does not think and act contrary to character since he is often manipulated to be rationally egoistic. His effective first-order desire to kill Ms. White conforms to his second-order desires. Plum's reasoning process exemplifies the various components of moderate reasons-responsiveness. He is receptive to the relevant pattern of reasons, and his reasoning process would have resulted in different choices in some situations in which the egoistic reasons were otherwise. At the same time, he is not exclusively rationally egoistic since he will typically regulate his behavior by moral reasons when the egoistic reasons are relatively weak—weaker than they are in the current situation (Pereboom 2001, pp. 112–13).

Pereboom's intuition is that Plum is not morally responsible for killing White in Case 1 because the neuroscientists' activities, which are beyond

Plum's control, determine his behavior. Further, Pereboom insists that Plum is not responsible even though his actions satisfy all compatibilist conditions of responsibility that leading compatibilist contenders set forth (p. 113). Pereboom then introduces a second scenario:

> **Case 2.** Plum is like an ordinary human being, except that he was created by neuroscientists, who, although they cannot control him directly, have programmed him to weigh reasons for action so that he is often but not exclusively rationally egoistic, with the result that in the circumstances in which he now finds himself, he is causally determined to undertake the moderately reasons-responsive process and to possess the set of first-and second-order desires that results in his killing Ms. White. He has the general ability to regulate his behavior by moral reasons, but in these circumstances, the egoistic reasons are very powerful, and accordingly he is causally determined to kill for these reasons. Nevertheless, he does not act because of an irresistible desire (Pereboom 2001, pp. 113–34).

Pereboom again believes that, although Plum satisfies compatibilist conditions, Plum is not morally responsible in Case 2 because the neuroscientists' programming, which is beyond Plum's control, determines his action (p. 114). Next, Pereboom advances a scenario in which the neuroscientists are replaced by parents, community, and other like candidates:

> **Case 3.** Plum is an ordinary human being, except that he was determined by the rigorous training practices of his home and community so that he is often but not exclusively rationally egoistic (exactly as egoistic as in Cases 1 and 2). His training took place at too early an age for him to have had the ability to prevent or alter the practices that determined his character. In his current circumstances, Plum is thereby caused to undertake the moderately-reasons-responsive process and to possess the first- and second-order desires that result in his killing White. He has the general ability to grasp, apply, and regulate his behavior by moral reasons, but in these circumstances, the egoistic reasons are very powerful, and hence the rigorous training practices of his upbringing deterministically result in his act of murder. Nevertheless, he does not act because of an irresistible desire (Pereboom 2001, p. 114).

Finally, Pereboom constructs a fourth case of ordinary upbringing in the context of causal determinism:

> **Case 4.** Physicalist determinism is true, and Plum is an ordinary human being, generated and raised under normal circumstances, who is often but not exclusively rationally egoistic (exactly as egoistic as in Cases 1–3). Plum's killing of White comes about as a result of his

> undertaking the moderately reasons-responsive process of deliberation, he exhibits the specified organization of first- and second-order desires, and he does not act because of an irresistible desire. He has the general ability to grasp, apply, and regulate his behavior by moral reasons, but in these circumstances the egoistic reasons are very powerful, and together with background circumstances they deterministically result in his act of murder (Pereboom 2001, p. 115).

As noted, Pereboom proposes that Plum is not morally responsible for killing White in Case 1. He also believes that there are no morally relevant differences that bear on responsibility between any two contiguous cases. Hence, Pereboom infers, Plum is not morally responsible in the last scenario describing ordinary upbringing against the backdrop of determinism. Because Plum or his action of murder satisfy compatibilist conditions of responsibility in each case, Pereboom further concludes that this sort of counterexample/generalization strategy undermines any compatibilist candidate. If Pereboom's conclusion is on the mark, there are strong reasons to be skeptical about the possibility of distinguishing, in a principled way, between cases involving CNC manipulation and those involving mere determination.

3.6. RESPONSE TO THE FOUR-CASE ARGUMENT

We believe that our relational account of authenticity enables us to differentiate, when such differentiation is called for, among the four cases. Construe Case 1 in any reasonable way in which it is *clear* that Plum is *not* morally responsible for killing White. So, for instance, imagine that the neuroscientists at the pre-initial scheme stage implant, at the appropriate time, suitable salient action-producing events that guarantee that Plum's controlled actions meet the goals of the scientists but that *undermine* epistemic or control requirements of responsibility. *Baseline-1,* with the pertinent facts, implies that the causal route to the acquisition of these elements is deviant. However, now consider Case 2 that is amenable to development in different ways. In one variation, the neuroscientists program Plum by implanting in him doxastic and pro-attitudinal elements that are constituents of his initial evaluative scheme and none of these elements undermine responsibility for later behavior by undermining epistemic or control requirements of responsibility. Here, the process of initial scheme acquisition is speeded up considerably, assuming this is possible. We suppose that when Plum kills White, his deadly deed causally stems from an authentic evolved scheme. The causal routes to the acquisition of Plum's initial scheme and his evolved scheme qualify as baseline (normal), respectively, according to *Baseline-1* and *Baseline-2*. We, thus, see no reason to deny that he is responsible under these circumstances.

In a second variation of Case 2, the implanted elements constitutive of Plum's initial evaluative scheme do subvert responsibility for later behavior, including the deadly deed, that issues from these elements. In this variant, the verdict that derives from *Baseline-1* is that Plum is not morally responsible for killing White.

In a third variation, just as in the first, the implanted initial scheme is authentic. Suppose, though, that pursuant to initial scheme acquisition, the sly neuroscientists program Plum to behave in various ways, including the following. They implant in Plum a set of relevant desires that become activated to exert influence on Plum's behavior under circumstances conducive to Plum's killing White, and they implant a set of suitable beliefs that come to Plum's mind in these circumstances. Imagine that Plum's reasoning to kill White issues from these beliefs and desires without engaging elements of Plum's authentic evolved scheme. Then, once again, *Baseline-2* generates the result that Plum is not morally responsible for killing White.

Pereboom claims that a compatibilist who takes Plum to be morally responsible in Case 3 must show how this case differs from Case 2. Suppose one believes that Plum is not morally responsible in Case 2 because one has either the second or third variation of this case in mind. Case 3 may differ from Case 2 in that it is akin to the first variant of Case 2. So, relative to the niceties of the pertinent cases, *Baseline-1* and *Baseline-2* allow us to draw a principled distinction between a scenario in which Plum is morally responsible and a scenario in which he is not morally responsible for the killing.[20]

In conclusion, we do not, of course, pretend to have given a complete account of authentic evaluative schemes and, hence, a comprehensive account of deviant or baseline causal routes to the acquisition of salient action-producing elements. We have made a start however. When manipulation subverts responsibility, it is because causal routes that are deviant relative to baselines have been effectively exploited. We believe that various criticisms of compatibilist or incompatibilist accounts of freedom-level, responsibility-grounding control that invoke manipulation may have the semblance of cogency but only because baselines are left unspecified. We have proposed that when baselines are taken seriously, many of these criticisms fall by the way.

In *Appendix A,* we summarize and evaluate various other approaches to handling the troubling quandary of manipulation. Assessing how it stacks up against prominent rivals constitutes an important, partial defense of our account. Recently, McKenna has developed an interesting response to Pereboom's four-case argument. If cogent, his response sheds doubt on our rejoinder to Pereboom. In *Appendix B,* we assess McKenna's engaging response.

In the next chapter, we defend the relational conception of authenticity against an argument that is aimed at its core: there is little, if any, reason to think that even a globally victimized post-surgery agent, such as Jenny, is not morally responsible for behavior that expresses her engineered-in actional springs.

4 Forward-Looking Authenticity in the Internalism/Externalism Debate

4.1. INTRODUCTION: THE MAGICAL AGENTS OBJECTION

Cases such as Psychohacker, involving manipulation that is intuitively responsibility undermining, largely (but not exclusively) motivate the authenticity requirement that we have been developing. We have submitted that post-surgery Jenny is not morally responsible for intentional actions that express, for instance, her implanted desires and beliefs. There is, however, a powerful objection against regarding an agent such as victimized Jenny as not responsible. This objection appeals to "magical agents": individuals very much like normal, healthy, adult human beings who spring into existence with evaluative schemes fully in place. Perhaps these instantaneous agents are the product of others; maybe they are chance accidents of conspiring natural forces. It is alleged that such agents can be morally praise- or blameworthy for at least some of their conduct. It is further alleged that if this is so, then it is also true that agents such as maltreated Jenny may well be morally responsible for behavior stemming from her engineered-in antecedents of action.

One principal goal of this chapter is to explain why this objection fails. Our response to the objection secures a second primary aim: it reveals our stance on an important debate—the internalism/externalism debate—in the literature on freedom and responsibility. Roughly, externalists (some may prefer the label "historicists") about freedom and responsibility believe that freedom and responsibility are essentially historical phenomena; whether agents are free or responsible vitally depends on, for example, *how* agents come to have the psychological features that they have. Internalists (some may prefer the label "structuralists") deny that the historical genesis of such features plays any pivotal role in correct ascriptions of freedom or responsibility. The view that we defend incorporates elements of both internalism and externalism and is, in this respect, hybrid.

4.2. FRANKFURT'S PARTICIPATION PRINCIPLE

Rosa is a magical agent: she was "born" an instant ago, and with the exception of her unconventional entrance into life, she enjoys the full complement

of features that autonomous or morally responsible agency demands. She hears about the plight of the children in Niger; whipping out her "magical wallet," choke full of large denomination bills, she makes a bountiful donation to a well-reputed, pertinent charity. Is she morally praiseworthy for this deed? Externalists about responsibility may be prone to say that owing to her lacking a past, and owing to responsibility's being an essentially historical phenomenon, she is not deserving of praise. Internalists who believe that the past plays no such heavy hand in ascriptions of responsibility will be inclined to judge otherwise.

We believe that Harry Frankfurt's *Participation Principle* is both useful in adjudicating this dispute and in shedding light on global manipulation scenarios.

> *Participation Principle:* A person is morally responsible for an action only if he is properly implicated (alternatively, "invested" or "engaged") in the action.

As it stands, the principle is somewhat amorphous but one gets a sense of its import as one traces its evolving incarnations in Frankfurt's penetrating discussions on free and responsible agency. So, for example, in "The Problem of Action," in which Frankfurt's chief concern is to argue against the view that causal theories of action provide a satisfactory account of the nature of action, Frankfurt writes:

> [These theories cannot] give any account whatever of the most salient differentiating characteristic of action: during the time a person is performing an action he is necessarily in touch with the movements of his body in a certain way, whereas he is necessarily not in touch with them in that way when movements of his body are occurring without his making them. (Frankfurt 1978/1988, p. 71)

In "Freedom of the Will and the Concept of a Person," Frankfurt (1971/1988) submits that wantons do not care about what (first-order) desires move them to action; such agents lack preferences regarding, for instance, which conflicting first-order desires are effective. Unlike *persons,* agents who have second-order volitions—second-order desires concerning which first-order desires should move them to action—wantons are not morally responsible for their actions; they are not suitably invested in them. For Frankfurt, to care about something is to be active. So, for example, he says, "with respect to those things whose importance to . . . [a person] derives from the fact that he cares about them, the person is necessarily active." If a person does not care about something which turns out to be important to him, "the person is passive with respect to the fact that the object is important to him." (Frankfurt 1992/1999, p. 87) Frankfurt adds that caring "presupposes both agency and self-consciousness. It is a matter

of being active in a certain way." (Frankfurt 1982/1988, p. 83) This same theme is echoed in "Autonomy, Necessity, and Love" when Frankfurt says that "insofar as a person's will is affected by considerations that are external to it, the person is being acted upon. To that extent, he is passive. The person is active, on the other hand, insofar as his will determines itself" (Frankfurt 1994/1999, p. 133). David Velleman crisply summarizes what the *Participation Principle* strives to encapsulate:

> What primarily interests Frankfurt . . . is the difference between cases in which a person "participates" in the operation of his will and cases in which he becomes "a helpless bystander to the forces that move him." And this distinction just is that between cases in which the person does and does not contribute to the production of his behaviour. (Velleman 1992, p. 470)[1]

Another enduring theme in Frankfurt is a particular understanding of in what participation or investment in an action consists. Such participation is a matter of activity on the agent's part that generates a set of first-order desires or attitudes she cares to have, desires internal to her "volitional structure" to which she decisively commits herself and with which she identifies. The unwilling addict, who shoots up despite identifying with the desire to refrain from taking the drug, is not morally responsible for indulging. Although taking the drug is an intentional action on her part, the unwilling addict is not invested in this action; she is "passive" with respect to it.

4.3. THE PARTICIPATION PRINCIPLE AND GLOBAL MANIPULATION CASES

Let us introduce another of Mele's illustrations of a global manipulation case. Ann and Beth are both philosophy professors but Ann is far more dedicated to the discipline. Wanting more production out of Beth and not scrupulous about how he gets it, the dean of the University enlists the help of new-wave neurologists who implant in easy-going Beth, Ann's hierarchy of values. Just as in Jenny's case, the implanted pro-attitudes are practically unsheddable: given Beth's psychological constitution, ridding herself of the attitudes is not a "psychologically genuine option" under any but extraordinary circumstances.[2] The global manipulation results in Beth's being, in germane respects, the psychological twin of Ann. The induction leaves unscathed values, beliefs, desires, and so forth which pre-manipulated Beth possessed and which can co-exist more or less harmoniously with the engineered-in pro-attitudes.[3] Upon completion of her transformation, is Beth morally responsible for her initial philosophical activity which expresses her unsheddable engineered-in values? Again, the thought experiment is effective only on the supposition, which we find no reason to reject, that the

intervention leaves personal identity intact: pre-manipulated Beth is identical to her post-manipulated later self.[4]

It is generally taken that global manipulation cases signal a divide between internalist and externalist positions on free action, moral responsibility, and autonomy. (To facilitate exposition, we shall henceforth primarily address moral responsibility with the implicit understanding that corresponding, suitably qualified things hold true of autonomy and free action as well.) To formulate these positions more perspicuously, we assume that at least certain elements of a person's psychology play an essential role in responsibility ascriptions. Specifically, we assume that various conditions of moral responsibility, such as agency and control (or freedom) conditions cannot be specified independently of invoking these elements; these conditions essentially appeal to these elements. Call such psychological elements "responsibility-grounding psychological elements" and call the conditions of responsibility that essentially appeal to these elements "psychology implicating conditions of responsibility." *Externalism* (on our view) is the thesis that the psychology implicating conditions of moral responsibility cannot be specified independently of facts about how the person acquired her responsibility-grounding psychological elements. The salient idea is that facts about one's history or past in the external world that bear on the acquisition of one's responsibility-grounding psychological elements are pertinent to whether one's actions are free and, hence, pertinent to whether one can be morally responsible for them. *Internalism* (as we understand it) is the thesis that the psychology implicating conditions of moral responsibility can be specified independently of facts about how the person acquired her responsibility-grounding psychological elements. As David Zimmerman (2003*a,* p. 642) comments, according to internalists, "the conditions of autonomous agency are limited to features internal to a person's attitude-system during the period of deliberation that proximally precedes action."

Mele takes global manipulation cases to support externalism:

> Ann, by hypothesis, is autonomous; but what about Beth? . . . By instilling new values in Beth and eliminating old ones, the brainwashers gave her life a new direction, one that clashes with the considered principles and values she possessed prior to manipulation. Beth's autonomy was violated, we naturally say. [Footnote omitted.] And it is difficult not to see her now, in light of all this, as heteronomous to a significant extent. If that perception is correct, then given the psychological similarities between the two agents, the difference in their current status regarding autonomy would seem to lie in how they *came* to have certain psychological features that they have, hence in something *external* to their here-and-now psychological constitutions. That is, the crucial difference is *historical;* autonomy is in some way history-bound. (Mele 1995, pp. 145–46)

It would not be out of the ordinary to expect a partisan of the *Participation Principle* to plump for a verdict that matches Mele's. For there seems to be a fairly transparent sense in which transformed Beth is not suitably *invested* in her initial, pertinent actions; these actions spring from causal antecedents that, at least intuitively, are "foreign" to Beth. Frankfurt, though, argues for the contrary verdict. We introduce salient passages in which Frankfurt expresses commitment to internalism:

> A manipulator may succeed, through his interventions, in providing a person not merely with particular feelings and thoughts but with a new character. That person is then morally responsible for the choices and the conduct to which having this character leads. We are inevitably fashioned and sustained, after all, by circumstances over which we have no control. The causes to which we are subject may also change us radically, without thereby bringing it about that we are not morally responsible agents. It is irrelevant whether those causes are operating by virtue of the natural forces that shape our environment or whether they operate through the deliberate manipulative designs of other human agents. (Frankfurt 2002, pp. 27–28)

> [T]o the extent that a person identifies with the springs of his actions, he takes responsibility for those actions and acquires moral responsibility for them; moreover, the questions of how the actions and his identifications with their springs are caused are irrelevant to the questions of whether he performs them freely or is morally responsible for performing them. (Frankfurt 1975/1988, p. 54)

What drives Frankfurt's somewhat surprising verdict on cases such as the Ann/Beth case given that the *Participation Principle* is well-entrenched in his works and that Beth seems to be on the sidelines regarding her relevant initial post-induction intentional behavior? One reason is faith in decisive wholehearted identification's being the correct currency to cash out the slippery notion of participation, investment, or activity.[5] In this chapter, we will not directly contribute to the vast literature on the strengths or weaknesses of whether appeal to appropriate hierarchies of desires or attitudes will, in the end, suffice to account for responsible agency. Some of the results of the ensuing inquiry, though, may shed indirect light on this on-going debate.[6] A second reason may be strong conviction in the truth of internalism. The obvious concern is that when global induction cases are invoked to nudge the philosophical opinion one way or another—in favor or against internalism—this reason appears to be question begging. This charge can be escaped, though, if one can provide independent motivation for the plausibility of internalism. In the next section, we explore a fascinating line of reasoning that, if cogent, would seem to supply the requisite autonomous support. The approach invokes the phenomenon of instantaneous or magical agents.

4.4. MAGICAL AGENTS AND GLOBAL MANIPULATION

Assume that the concept of *instantaneous autonomous agency* is coherent and, hence, that Rosa *could* have entered life an instant ago, bearing all the features that responsibility requires.[7] The possibility of instant autonomous agents may be conscripted in an argument against Mele's verdict (which matches ours) concerning responsibility on Beth in the Ann/Beth case. If sound, this argument provides support for internalism insofar as it removes what many have taken to be a major card in favor of externalism.

An instructive incarnation of such an argument is McKenna's. McKenna's chief exhibit is magical agent Suzie Instant who comes into existence at an instant as a psychologically healthy woman much like "any other normally functioning thirty-year old person." (McKenna 2004*a*, p. 180) The handy work of a God, Suzie has a complement of false beliefs "according to which she has lived a normal human life for thirty years" and a range of "values and principles that are unsheddable." (McKenna 2004*a,* p. 180) Suzie (falsely) believes that she has acquired her values through sustained effort over years leading up to what she thinks is her thirtieth birthday. She takes pride in this belief, all the while thinking that she is responsible for her efforts which she has freely exerted. We are to assume, in addition, that Suzie is a richly self-controlled person who is able to resist the inclination to act from weakness of will and that she "satisfies something like Frankfurt's hierarchical account of freely willed conduct . . . [and] is reasons-responsive. . . . In short, Suzie satisfies the juiciest nonhistorical demands [a compatibilist might venture]" (McKenna 2004*a,* p. 180). Finally, we are to suppose that Suzie has a robust and relatively consistent range of (false) beliefs about her history similar to those that any psychologically healthy person would have. McKenna invites us to ponder the following:

> [S]uppose . . . Suzie is presented with the option to do . . . A or B. . . . [B] involves a violation of a value that is unsheddable for her. . . . [A] involves acting from one of her unsheddable values. Suzie A-s, acting as her unsheddable value counsels. Supposing that compatibilism is true, it is not clear to me that Suzie did not act freely or responsibly. I can't see how a causal history that zeroed in on this Suzie all in an instance . . . renders . . . [her] unfree in a way that she would not be if instead some causal history or other unfolded over the course of thirty years. Note that . . . when Suzie A-ed from her unsheddable value, she was *not* compelled to do so. Her doing so was *nothing like* acting upon an irresistible desire. It would be natural to say that she A-ed freely—in at least some non-question begging, restricted sense of freely. . . . To press the point, suppose that every now and then this same god who created Suzie Instant visits another possible world and there creates another thirty year old Suzie, Suzie Normal, in the normal zygote manner. Other times

> she creates a (seemingly) thirty year old Suzie, Suzie Instant, at an instant. Now suppose that Suzie Normal at the age of thirty arrives at the precise point where she comes to be a historical duplicate of Suzie Instant. Suzie Normal faces the exact same choice between options A and B as Suzie Instant faces. Just like Suzie Instant, Suzie Normal opts to do A. . . . The crunch is now upon us: How is it that Suzie Instant is rendered not free and morally responsible when she A-s at the relevant time merely by virtue of the fact that the causal history giving rise to her action came compressed in a momentary package where Suzie Normal's history chugged along over the course of thirty years? A difference here seems arbitrary. (McKenna 2004*a*, pp. 180–81)

To secure the conclusion that victimized Beth may well be morally responsible, McKenna continues:

> Suppose that it is arbitrary to claim that Suzie Instant is not free and responsible but Suzie Normal is. Here we have a case and some attendant intuitions speaking on behalf of a nonhistorical conclusion as regards the case of Suzie Instant. But of course, intuitions about varying cases can compete. So it is time for Suzie Instant, Suzie Normal, Ann, and Beth to meet. Let's start with Suzie Instant and Ann, and let us stick with a case in which each A-s as opposed to B-s in such a way that their respective acts of A-ing issue from their respective unsheddable values. Recall, unlike Beth, Ann acquired her wonderful professorial values under her own steam, with the sort of history that Mele finds to be freedom and responsibility conferring. Suppose that by mere cosmic accident, not even by the intentional design of the god who brought Suzie Instant into existence, Suzie Instant is a nonhistorical qualitative duplicate of Ann. She is so right down to her (false) beliefs about her history. If Ann recalls the hours of labor she spent knocking out her last article, Suzie Instant (falsely) recalls the hours of labor that she thinks she spent. Her psychic life, her memory of how she came to be is just as Ann's is. I submit that if Ann and Suzie Instant behaved in the same ways in the same circumstances, Suzie Instant's conduct should be regarded as free and responsible if Ann's is. If this result seems dubious, just start one step away from this. Make the cosmic accident that this god created Suzie Normal to be a qualitative duplicate of Ann, living out the very same life and history as Ann right up to the moment when Ann A-s instead of B-s. . . . But now, as I have argued above, we should treat the case of Suzie Instant no differently than we treat the case of Suzie Normal. Hence, we should treat Suzie Instant's case no differently than we treat Ann's. (McKenna 2004*a*, pp. 181–82)

McKenna (with some qualifications to which we shall return in Section 4.6 below) proposes that Beth is not relevantly different from Suzie Instant.

So if Suzie Instant is responsible for *A*-ing, Beth should be responsible for *A*-ing as well. We summarize the argument in this way:

1. Suzie Instant is morally responsible for *A*-ing. (This is because Suzie Normal is morally responsible for *A*-ing, and Suzie Instant is not pertinently different in the constellation of features required for moral responsibility from Suzie Normal.)
2. If 1, then globally manipulated Beth is morally responsible for *A*-ing. (This premise rests on the view that Beth is not relevantly different in instantiating the features responsibility requires from Suzie Instant.)
3. Therefore, globally manipulated Beth is morally responsible for *A*-ing.

4.5. WHY THE ARGUMENT FAILS

We accept the rationale for the first but not the second premise. The burden of this section is to explain why this is so. We first attend to a possible problem that infects McKenna's test case. Barring special reasons to believe otherwise, internalists have no reason to resist the customary view that non-culpable ignorance is (or may well be) an excusing condition. This saddles Suzie's Instant's tale with an instant glitch: though the lion's share of her beliefs are false, Suzie is *not* culpably ignorant regarding them. To handle this concern, perhaps the case can be suitably tweaked so that when Suzie *A*-s, her *A*-ing does not implicate false beliefs; thus, her *A*-ing evades worries concerning satisfaction of epistemic constraints on responsibility. Assume that the case can be so modified.

To understand why Premise 2 is on slippery footing, we redirect attention to our distinction between two stages in an individual's development (in the case of beings, like us, who acquire responsibility-grounding psychological elements over time). These are the stage prior to which one is a morally normative agent—an agent who is a suitable candidate for responsibility ascriptions—and the stage when or after one satisfies the agency requirements of responsibility.[8]

Global manipulation cases, as standardly presented, such as the Ann/Beth case, raise concerns not about pre- but about fully-formed normative agents with evolved evaluative schemes. Beth, in Mele's example, is such an agent. Global manipulation has the effect of subverting normative agency, and thus, more generally, affecting agency, in this way: various pro-attitudinal and doxastic components of the individual's evaluative scheme are "replaced" by a different set. The replacement is not accomplished under the individual's own steam but occurs as a result of some process that totally bypasses the agent's capacities of deliberative control.[9]

Frankfurt's *Participation Principle* says that a person is morally responsible for an action only if he is properly implicated (alternatively, "invested"

or "engaged") in the action. We propose that the deep insight this principle captures may be expressed in this fashion:

> *an agent is suitably "in touch" with an action of hers—is properly "invested" in that action—only if the action causally stems from elements of an evaluative scheme of hers that is authentic.*

Whereas Frankfurt's understanding of agent participation appeals to decisive wholehearted identification, we understand agent investment as essentially associated with behavior causally deriving from authentic evaluative schemes. On our view, the evaluative scheme definitive of normative agency with which Beth finds herself after global manipulation (or a substantial part of it) is not authentic. Hence, when she *A*-s, and her *A*-ing expresses engineered-in values, she is not suitably invested in that instance of her *A*-ing.

What of Suzie Instant, though? Does her pertinent action—her *A*-ing—issue from an authentic evaluative scheme? Suzie Instant is, in certain respects, just like a young child whose evaluative scheme has not been acquired: others (or at least something in perturbations of the original Suzie Instant scenario) contribute to the child's or to Suzie's acquisition of an *initial* evaluative scheme. Depending on how the parable is expounded, Suzie Instant's initial evaluative scheme, just like that of a young child, qualifies as relationally authentic: we may assume that its pro-attitudinal and doxastic elements do not subvert responsibility for intentional behavior that has a subset of these elements as actional antecedents.

The first premise of McKenna's argument—that Suzie Instant is morally responsible for *A*-ing—hinges on the rationale that insofar as the features that ground moral responsibility are concerned, Suzie Instant is no different than Suzie Normal. We may grant this premise because we assume that when each of these agents *A*-s, her *A*-ing issues from an evaluative scheme that is authentic; Suzie Instant's initial scheme is authentic as is Suzie Normal's evolved scheme.

The second premise—if Suzie Instant is morally responsible for *A*-ing, then globally manipulated Beth, too, is morally responsible for *A*-ing—turns on the proposal that Beth is not relevantly different in instantiating the features responsibility requires from Suzie Instant (or Suzie Normal). The proposal is false. Beth's *A*-ing issues from components of an evaluative scheme which is inauthentic, whereas Suzie Instant's issues from components of an authentic scheme. In Suzie Instant's case, the concern is whether an *initial* evaluative scheme is relationally authentic; in manipulated Beth's case, the concern is whether substantial elements that replace various elements of a prior scheme are authentic. Again, an implication of our analysis is that not all global manipulation cases subvert responsibility on the assumption that a case, such as Suzie Instant's, passes as a variant of a global manipulation case, and that there is a principled way, one sensitive to whether what is engineered-in is an initial scheme or "replacement

elements" of a cluster of elements of a scheme already in place, to distinguish among such cases.

4.6. INTERNALISM'S DOMAIN

Turning, now, to an objection, one might well wonder whether this sort of differentiation of global manipulation cases merely begs the question against internalists. Internalists, roughly, insist that facts in the external world concerning how one's springs of action are acquired make no difference to moral responsibility. Siding with Mele's verdict on the Ann/Beth case, though, we have assumed that whether Beth's acquisition of (substantial) parts of her evaluative scheme is accomplished under her own steam *does* bear on responsibility. Hence, the complaint of question begging is motivated. There are considerations, however, against the legitimacy of this complaint.

First, a preliminary comment is appropriate. If one takes the *Participation Principle* seriously, it *should* matter how one's evaluative scheme is acquired. For if one is *altogether* divorced from the acquisition or modification of an evaluative scheme—and this is especially obvious once elements of an evaluative scheme have been acquired—how can one be appropriately invested in actions that causally issue from components of the scheme?

One might complain, though, that this merely reasserts the view that externalism is true without advancing anything new in its support. Frankfurt, for example, could simply reply, "Here's how one can be suitably invested: one *identifies* oneself with the springs of one's action, and thereby *takes* responsibility for it, regardless of the causal source or history of the acquisition of those springs." Indeed, commenting on a case in which a "Devil/neurologist" (D/n) manipulates a victim by providing the victim with a stable character or higher-order program—a set of rules—that, from the time of being instilled, determines the victim's mental and physical responses to his outer and inner environment without further intervention by the D/n, Frankfurt gives this very sort of response. He says that the victim may become autonomous "in the same way [non-manipulated, normal] others do: by identifying himself with some of his own second-order desires" (Frankfurt 1975/1988, p. 53). "Passive" second-order desires that the D/n covertly instills become "active" second-order volitions through a process of (presumably third-order) identification: "In virtue of a person's identification of himself with one of his own second-order desires, that desire becomes a second-order volition" (Frankfurt 1975/1988, p. 53).

However, this reply seems at loggerheads with the *Participation Principle*. What it overlooks is that the D/n may orchestrate the very process of identification itself: whenever the suitably programmed victim critically reflects upon such-and-such second-order desires with which he happens to find himself, a pre-installed higher-order rule becomes operational to the

effect that he automatically identifies himself with these desires and thereby turns them into second-order volitions. It is true that an identification with a second-order desire is itself a mental act of the third-order, or assume that it is. Yet, given Frankfurt's naturalized conception of autonomy, there is no reason whatsoever why third-order mental acts themselves could not be subject to what externalists regard as illegitimate causal influence. If identification is nothing more than a natural process, then no matter how complex and no matter what order of desires are at issue, it too can be covertly controlled.[10] Should the process itself be contrived, then it is not clear what sense is to be made of the claim that an agent can be appropriately implicated or invested in an action *by virtue of identifying herself* with some of its causal springs. With contrived identification, the agent is on the sidelines and is not properly engaged in the action in a manner in which he would have been were the identification not contrived. Actions which issue from desires with which the agent is *made* to identify are exemplars of actions which are produced in a way that bypasses the agent's capacities of control over her mental life.

Frankfurt's response to when one's desire is truly one's own (and thus, his response to when one is properly "invested" in an action that causally and non-deviantly derives from that desire) is straightforward: when one identifies oneself with that desire. The response to when *identification itself* is truly one's own, thus being identification that is not contrived and which, thereby, includes pertinent "participation" of the agent, cannot be that one is (somehow) "identified" with the very process of identification. It is a myth to believe that identification can remain insulated from things like covert manipulation, and so can remain pristine, no matter what the history of the child's acquisition of pro-attitudes, deliberative principles, values, and so forth. Identification, no less than reasoning, *cannot* be divorced from one's deliberative principles; the notion of such "principle-independent identification" strikes us as incoherent. If such deliberative principles can be tainted in the manner in which we have explained—they may not be relationally authentic at the pre-initial scheme stage—identification, at a time when the normative agent into whom the child has developed can engage in identification, *inherits* this taint.

Second, speaking more directly to the charge of question begging, reflection on how initial evaluative schemes can be authentic, in particular, appreciation of the fact that with respect to the doxastic and pro-attitudinal components of children who are not fully developed normative agents, there is nothing like authenticity of these components *per se,* but only relational authenticity, one can isolate the domain of internalism. We recorded that instantaneous agents are like children in the respect that their evaluative schemes are initial and not evolved schemes. Just as it would not affect the *authenticity* of, say, a set of values that were implanted in a child at the pre-initial scheme stage that the implantation were accomplished by harshly paternalist means as long as the implanted values, or their means

of implantation, did not compromise later responsibility for behavior issuing from these elements, so it would not matter whether Suzie Instant's initial scheme were the product of deliberate engineering or something created *ex nihilo,* again, provided that elements of this scheme would leave unblemished responsibility for behavior that arises from these elements. (Of course, in cases of this sort, one need not deny that moral wrong has befallen the child, if such is so, or that Suzie has been unfairly treated.)

Focus on a case of instantaneous agency in which we have creation *ex nihilo* of a Suzie Instant type of agent—Suzie*. Suzie* has no history or past. Therefore, it should come as no surprise that facts about her history in the external world cannot have a bearing on the acquisition of her (fully-formed) evaluative scheme with which she is "born" and so cannot, in one way or another, affect responsibility for behavior that issues from elements of this initial scheme. If it is sufficient (other conditions assumed) for a position to qualify as internalist that facts about the agent's history in the external world have no bearing on autonomous or responsible agency, then magical agents vindicate internalism. It would be a mistake, however, to conclude from this that past facts should not have a bearing on responsibility with agents (like us) who *do* have a past.

To elaborate, suppose that moments after her "birth," Suzie* falls victim to Ann/Beth style induction. We see little reason why internalists should deny that the manner in which Suzie* acquires the "new" components of her evaluative scheme has distinct implications for responsibility even though pre- and post-manipulated Suzie* are equipped with the psychology that both internalists and externalists would deem sufficient for responsibility but for its odd provenance. We registered previously that moral responsibility has several requirements. Setting aside the demands of agency and the dispute over whether responsibility has a strong historical dimension, both internalists and externalists can agree on these other requirements—such as epistemic and control constraints—and can agree on what factors are responsibility-subverting factors. Hence, internalists can acknowledge that factors such as coercion that may affect freedom, or factors such as deception that may compromise epistemic requirements of responsibility, may well be responsibility subversive. Then there is no reason for internalists to resist the view that factors that compromise agency requirements of responsibility—factors, for instance, that threaten moral normative agency—may also be responsibility subversive. In the initial scenario involving her "birth," there is no question about subversion of agency; actions that causally arise from Suzie*'s authentic initial evaluative scheme satisfy the *Participation Principle*. In the latter scenario involving global manipulation, there is an obvious concern with the subversion of agency; actions that causally arise from Suzie*'s modified evaluative scheme fail to respect the *Participation Principle*.

If internalists grant that factors concerning normative agency can affect responsibility and they grant that there is no reason to deny that, with

agents who have pasts, historical factors can affect responsibility, there is every incentive for internalists who welcome the *Participation Principle* to embrace our verdict that manipulated Beth is not responsible for her pertinent actions. They can do all of this consistently with holding on to internalism: magical agents are a conceptual possibility; such agents have no pasts. Thus, if such agents are morally responsible, it is not a requirement of responsibility that one have a past.

In summary, assume, though this is somewhat stretched, that Suzie*'s initial scenario (in which she springs into existence) qualifies as a global manipulation one (a case in which a deity created her would be more apt) as does the latter scenario. In the latter scenario (in which Suzie* falls victim to Ann/Beth style manipulation) unlike in the former, we have proposed that internalists need not disagree that facts about Suzie*'s history in the external world—facts regarding how Suzie* acquires components of her evaluative scheme—do have a bearing on responsibility. We can give a principled account of why internalism "holds" in the former but not in the latter scenario. Hence, one can side with our verdict that Beth is not morally responsible by calling upon the *Participation Principle* without any question begging against internalism.

Still, perhaps an internalist will object that what *makes* manipulated Beth not responsible is not her past but her present. What happened in the past might have produced the conditions that undermine her agency, and thus her responsibility now, but what renders her not responsible is solely a matter of the current structure of her will. Indeed, the internalist and externalist may actually agree on the responsibility verdict for all cases, but their dispute is over the *criterion* of responsibility—what it is that makes someone responsible or not—and *that* dispute will remain intact through all such agreement on specific cases. Simply to assert otherwise, though, is still question begging.

First, we respond directly to the objection. We then advance further grounds to support the contention that our strategy is not question begging. The objection presupposes that what renders manipulated Beth *not* responsible is "solely a matter of the current structure of her will." However, this is a *non-sequitur* if Beth, just like Suzie Instant, "satisfies the juiciest nonhistorical demands" an internalist might advance. (McKenna discerns no relevant difference among Suzie Instant, Suzie Normal, and victimized Beth, asserting that each *is* morally responsible.) Maybe the thought is that if manipulated Beth (or her psychology) satisfies the *correct* internalist criterion of responsibility—whatever it may be—then manipulated Beth *is* responsible for her post-transformation acts. This, though, flies in the face of the objector's claims that Beth is *not* responsible and that the internalist and externalist "may actually agree on the responsibility verdict for all cases."

In the context of this dispute between externalists and internalists, charges of question begging are delicate. To safeguard begging the question against

one or the other of these factions, it is helpful to assume, to the extent that this is possible, the stance of theorists who have no *pre-commitments* to either internalism or externalism but who take seriously the following riddle: it is intuitively plausible that manipulated Beth is not morally responsible. It is intuitively plausible that Suzie Instant is morally responsible. What accounts for this asymmetry in intuitions that, in turn, may shed light on the credentials of the seemingly competing positions?

Adopting this sort of neutral stance, we have *not* argued in this fashion: "the initial intuition that Suzie Instant is morally responsible is nonproprietary because, *ab initio,* it stacks the deck against the externalist." We do not think that this sort of inauguratory strike is of any help in resolving (or dissolving) the riddle anymore than would be the parry that Suzie Instant is morally responsible because she satisfies the juiciest internalist nonhistorical demands. Nor have we argued in this manner: "There is no plausible way for an externalist to explain why Suzie Instant is morally responsible consistently with the externalist's explanation of why victimized Beth is not morally responsible. So it must follow that Suzie Instant is not morally responsible." This thread of reasoning is, surely, given the dialectical context, unacceptable.

Rather, we have first *conceded* that Suzie Instant is morally responsible. This is an important plank in our defense against the charge of question begging. We have then tackled head-on the challenge implicit in McKenna's paper: Suzie Instant has no past. So facts in the external world in her past can have no bearing on why, if she is morally responsible, she is so. Hence, externalists who believe that, necessarily, facts in the external world in an agent's past can have a pronounced bearing on responsible agency cannot maintain the verdict that Suzie Instant is morally responsible consistently with their explanation of why manipulated Beth is not morally responsible. We have picked up the gauntlet. We have proposed a compromise of sorts between internalism and externalism: moral responsibility does not require that one have a past but it does require that one not have certain kinds of past. The relational view of the authenticity of an agent's initial evaluative scheme plays a fundamental role in securing the first half of this hybrid view.

Finally, any comprehensive theory of responsible agency should not shun the enormously complex issue of explaining how the child who begins life as an individual who is not morally responsible for any of her conduct (or who is non-autonomous), eventually turns into an agent who can be to praise or blame for various actions of hers (or who is autonomous). We have provided a sketch of certain ingredients of the story, and we have done so *without* presupposing internalism or externalism. Both internalists and externalists are free to co-opt what they see of value in our contribution.

Third, McKenna realizes that an externalist may attempt to turn the tables on the internalist by insisting that there are powerful intuitive grounds to believe that globally manipulated Beth is not morally responsible

for *A*-ing, and hence, that if Beth is relevantly like Suzie Instant (something we have questioned), then Suzie Instant, too, should not be morally responsible for *A*-ing. The internalist, though, may insist that the Ann/Beth case does beg the question against the internalist because manipulated Beth is not pertinently different from Suzie Instant who *is* morally responsible for *A*-ing. McKenna fears that the dialectic, at this stage, could "quickly degenerate into a stalemate with no resources for settling the matter beyond the tug of competing intuitions elicited by competing examples" (2004*a,* p. 182). It will be enlightening to assess McKenna's claim that "further considerations on behalf of the nonhistorical compatibilist like Frankfurt . . . might help pull the case of [manipulated] Beth into the nonhistorical camp" (McKenna 2004*a,* p. 182).

The first "further" consideration enjoins us to keep firmly in our minds that Beth satisfies the "very richest of compatibilist-friendly non-historical properties" and that we would respond to victimized Beth or Suzie Instant if we were to have a "moral transaction with one of them" in just the way in which we would respond to agents who are unquestionably responsible (McKenna 2004*a,* pp. 182–83). However this consideration cuts no ice at all: if one were cognizant of Beth's history or of Suzie Instant's peculiar origins, it is *not* in the least obvious that one would respond to these agents in the manner in which McKenna proposes that one probably would.

A second consideration seeks to remind us that there are "more moral judgments to go around than those that have to do just with the relevant action figuring in the manipulation case at issue" (McKenna 2004*a,* p. 183). So, for instance, McKenna submits that whereas Ann is morally responsible for coming to have the unsheddable values and, presumably, certain other features of the character that she has, Beth and Suzie Instant are not responsible for these things. So although the internalist "can argue that Ann, Beth, and Suzie Instant are all equally free and equally morally responsible *with respect to their acts of A-ing* . . . Ann is free and morally responsible for more than what Beth and Suzie Instant are free and morally responsible for" (McKenna 2004*a,* p. 183). McKenna suggests that what contributes to explaining "away the counterintuitive appearance of the judgment that Beth is free and morally responsible for *A*-ing is a failure to give sufficient attention to this fact" (McKenna 2004*a,* p. 182). In addition, McKenna claims that whereas the manipulators wronged Beth and violated some of her rights, nothing of this sort is true of Ann or Suzie Instant. He writes:

> Perhaps part of our reluctance to treat Beth as freely and responsibly A-ing is that we wrongly think that in making such a judgment, we are not recognizing the quite clear violations of Beth's rights as a person. But we can recognize that Beth freely and responsibly A-ed and *still* draw appropriate moral judgments about the moral wrongs done to Beth and how she deserves to be treated in light of that history. What, we might ask, could count as a proper moral response to Beth

> for her having suffered from someone else deciding for her what kind of person she should be? This question can be given a rich answer even if, now that she is this different (sort of) person, we are warranted in thinking that she is a person who acts freely and responsibly for what she now does. (McKenna 2004*a,* p. 184, note omitted)

Again, these considerations do little, if anything, to break the "stalemate" that McKenna describes. All parties to the dispute concerning whether Beth is morally responsible may well agree that certain moral judgments or assessments true of manipulated Beth need not be true of Ann. Still, they may well disagree on whether Beth is morally responsible, the disagreement stemming from a disagreement about whether the origin of the agent's actional springs has a bearing on responsibility. Indeed, it would be methodologically appropriate, in a dispute of this sort, to guard against muddying the waters by failing to keep squarely in mind the truism that extraneous factors, such as certain moral assessments other than ones having to do with responsibility, may differentially apply.

Finally, appealing to an account of blameworthiness, McKenna proposes that features about the nature of responsibility tell against the view that manipulated Beth is not morally blameworthy. It is beyond the ken of this paper to asses McKenna's analysis of blame- and praiseworthiness. The significance of this consideration to which we wish to draw attention is the following. McKenna's appeal to the nature of blameworthiness suggests that the verdict on the Ann/Beth case is to be settled, in part, by drawing on an entire theory of responsibility. We applaud this strategy. We simply wish to emphasize that the historicists that McKenna targets, such as Mele, Fischer, and Haji (though Haji turns out to be a "hybridist"), also embed conclusions concerning the influence of facts in the external world on responsibility in theories of moral responsibility.[11] The *Participation Principle* itself plays a critical, guiding role in Frankfurt's own theorizing about freedom, responsibility, and autonomous agency. The final "proof," then, lies in the pudding: which of the theories is superior, a theory that "validates" externalism (or at least "limited externalism": externalism with respect to agents who have pasts) or one that implies the truth of internalism?

4.7. FURTHER OBJECTIONS AND RESPONSES

We now respond to two further objections to our hybrid position. The first concerns actions for which an agent is allegedly morally responsible but which express changes in outlook which seemingly occur as a result of bypassing the agent's minimal capacities of reflective control. The actions in question may, for example, be expressive of certain "new values," the acquisition of which is not under the agent's control. We are invited to think of a modification of Frankfurt's "volitional necessity" cases, in which, say, many of one's

values are changed by an overwhelming care or commitment one has to some object. The phenomenon is defined as a case in which one cannot do otherwise, in a very important sense, and, further, as one in which one would not *want* to do otherwise, given that this would involve a betrayal of something about which one deeply cares. So we can imagine one's values changing in response to some overwhelming care (as in loving someone), and even though one cannot prevent the change, one would not *want* to prevent it. It is far from settled whether such cases involve the agent's "exercising" her initial scheme in any active sense. In addition, there is this sort of case: suppose that a selfish person's evaluative scheme undergoes sudden and drastic change through her witnessing some catastrophe—the devastation caused by the 2005 earthquake in Pakistan, say—and that the nature of her actions is accordingly drastically changed. We would be reluctant to declare her not morally responsible for these actions, even though the change in her outlook seems not to have been carried out under her own steam.

In response, taking the "necessities of love" as paradigm instances of volitional necessity, according to Frankfurt, it appears that the volitional necessity to which a lover is subject involves his being irresistibly *motivated* to act in the interest of his beloved and his irresistibly *identifying* with this motivation:

> The lover cannot help being selflessly devoted to his beloved. . . . It may seem that in this respect love does not differ significantly from a variety of other familiar conditions. There are numerous emotions and impulses by which people are at times gripped so forcefully and moved so powerfully that they are unable to subdue or resist them. . . . But irresistible forces do not invariably oppose or conflict with desires or intentions by which we would prefer to be moved. They may move us irresistibly precisely in ways that we are wholeheartedly pleased to endorse. There may be no discrepancy between what we must do and how we would in any event wish to behave. In that case, the irresistible force is not alien to us at all. (Frankfurt 1994/1999, pp. 135–37)

Recall our gloss on one's evaluative scheme's being engaged or exercised (in Chapter 3, Section 3.3). We have various capacities of deliberative control. The having of these capacities supervenes upon features or constituents of one's evaluative scheme. An agent's evaluative scheme is not engaged in, for instance, acquiring a desire, if the acquisition of that desire bypasses all of the agent's capacities of deliberative control. Presumably, the motivation that the lover acquires to act in his beloved's interest is *not* motivation that fails to engage the evaluative scheme of the lover. And the lover's identifying with this motivation is, again, presumably, something that does not bypass the lover's capacities of deliberative control since, as we noted, identification cannot be divorced from one's deliberative principles.[12]

As for cases involving sudden or drastic conversions, as Mele argued, whether the agent *is* morally responsible for the pertinent conduct piv-

ots vitally on the filling in of relevant details.[13] On the one hand, suppose (implausibly) that the witness to the earthquake undergoes the changes that she does because, at the time of the devastation, God implants in her a powerful disposition to be charitable, the implantation mirroring the implantation of manipulated Beth's new values. In this case, it is less than transparent whether the witness *is* morally responsible for her pertinent deed. On the other hand, imagine that the traumatic event generates in the agent an insight into the human condition which she then evaluates, and which subsequently moves her to change her outlook. In this variation of the scenario, it is false that the agent's evaluative scheme is idle: her exercising her capacities of deliberative control is crucial in explaining the change.

The second objection concerns alleged equivocation on the term "autonomy." It may be put to us that the reason we think Beth's autonomy has been compromised is that the manipulator has indeed violated her autonomy in the *moral* sense of the word—she gave no *consent* to the interference—but he did not necessarily violate her autonomy in the *responsibility-providing* sense of the word. So we can certainly agree with Mele that Beth's moral autonomy has been undermined, but if we infer from our agreement on that *term* that her responsibility was then undermined, we may very well be making a mistake: her failure to consent to the interference does not render her non-autonomous in the sense (ostensibly) required for moral responsibility.[14]

Like Mele, we take the global induction of the sort exemplified in Beth's case to be responsibility undermining but the objector (perhaps an internalist) does not. Presumably, though, the objector regards *some* forms of manipulation or treatment, or *some* sorts of interference, as responsibility subverting. Assume that Hal, in the absence of his consent, has been subjected to this sort of treatment. Assume, further, that some externalists do not regard this type of treatment as threatening responsibility. It would be ineffective for externalists of this bent to argue against an internalist dissenter in this way: "As Hal did not consent to the treatment, his moral autonomy has been violated. But it is a mistake to infer from this that Hal is not responsible for his pertinent behavior." Our candidate internalist would presumably not be moved by this sort of argument. She may well agree with the imagined externalist that the treatment violates Hal's moral autonomy. Further considerations would be required to persuade her, but contrary to what she believes that the treatment itself is not responsibility undermining and hence that Hal is responsible for his pertinent behavior.

Reverting to Beth's case, it is open to the objector to supplement the "no consent argument" with additional factors that tell against Beth's not being morally responsible. For instance, the objector might appeal to "sudden conversion cases" or "sudden change of outlook" cases to convince us that in these cases, despite undergoing the sudden changes, the agents are still morally responsible for their germane behavior, and then add that Beth's case is not relevantly different from these cases. Our response to the first objection, though, casts doubt on whether the sudden change cases

can turn the trick. When they are interpreted in a way in which they are analogous to Beth's case, it is contentious whether the germane agents are morally responsible for their pertinent acts. When interpreted differently in the manner suggested, the germane agents may well be responsible.

Assume that all parties to the dispute agree that Beth's moral autonomy has been violated—Beth did not consent to being globally altered. The externalist might then call attention to a case involving deception. Tom deliberately misleads Jerry with a view to getting Jerry to perform certain deeds that would benefit Tom. The deception is successful: Jerry acts on the basis of these false beliefs. Assume that the deception is of the sort that internalists regard as responsibility undermining. Imagine that Jerry*, a counterpart of Jerry's, non-culpably acquires beliefs type- or near type-identical to the ones Jerry is misled into acquiring and that he is otherwise as similar as possible, in psychological profile, to Jerry. It may well be that Jerry* is morally responsible for the actions that (partly) causally issue from these beliefs even though Jerry is not. To account for the difference in responsibility ascriptions, a theorist may propose that both internalists and externalists should give serious consideration to the suggestion that it matters how one acquires the beliefs. An internalist may rejoin that if we draw on historical considerations, we will be forced to admit that in all sudden conversion or sudden change cases, the agent is not morally responsible for her pertinent behavior. However as we have stressed, this is not so. It is, thus, worth paying close attention to history. We do not claim to have secured *decisively* our hybrid view that responsibility does not require that an agent have a past but it does require that the agent not have a past of certain sorts. Though we have made concessions to the internalist, as the first clause of the hybrid view makes clear, we are not willing to admit that the internalist has won the day.

To wrap up, we started with a somewhat tenuous rendering of the *Participation Principle* which we believe captures an important insight. We agree that to be morally responsible for an action, its agent must be invested in the action. This principle, we proposed, strongly suggests that victims of global manipulation, such as Beth, are not morally responsible for their pertinent actions; they are on the sidelines with respect to these actions. In the course of our discussion, we presented a refined rendition of this principle: an agent is suitably invested in an action if that action appropriately issues from an evaluative scheme of the agent that is authentic. We argued that on this reading, the *Participation Principle does* support the verdict that globally manipulated victims, such as Beth, are not morally responsible for their germane actions, despite pressure to judge otherwise that the phenomenon of instantaneous or magical agency exerts. To argue for this conclusion, we theorized that it is crucial to distinguish between two stages in a person's history: the stage prior to which the person is a normative agent—an agent with an evaluative scheme who satisfies the agency requirements of responsibility—roughly, the stage of childhood, and the stage after which a person

is such an agent. Children, in the maturation process, acquire an initial evaluative scheme over time. Instantaneous agents—magical agents—are like children in that the evaluative scheme with which they are equipped are initial ones. Components of one's evaluative scheme must be truly one's own (or, in our terminology, "authentic") if one is to be morally responsible for behavior that issues from them. We suggested that at the pre-evaluative scheme stage (or with instantaneous agents), there is no authenticity *per se* of the doxastic or pro-attitudinal constituents of agents' evaluative schemes but only relational authenticity: springs of action are authentic insofar as they do not compromise the agent's being morally responsible, at future times when the agent satisfies the agency requirements of responsibility, for behavior that issues from these springs. The relational account of authenticity is instrumental in carving out the appropriate domain of internalism, consistently with maintaining the *Participation Principle's* implication that globally manipulated agents, such as Beth, are not morally responsible for their pertinent behavior.

In the next chapter, we extend our defense of our relational conception of authenticity by showing how it enables us to make inroads into two deep puzzles in the philosophy of education.

5 Authentic Education, Indoctrination, and Moral Responsibility

5.1. INTRODUCTION: BRIDGING THE METAPHYSICS OF RESPONSIBILITY AND PHILOSOPHY OF EDUCATION

We have proposed that there is a requirement of authenticity for moral responsibility. We have developed a relational account of authenticity, traced its implications for cases involving manipulation, compared these implications with the pertinent implications of prominent rival approaches (in Appendix A), and defused (in the previous chapter) an argument that directly questions whether authenticity is a condition on responsibility. These endeavors should go some way toward allaying those skeptical of responsibility's having any such requirement that the requirement is a bona fide one. We argued that our relational account generates intuitively satisfactory results in a wide range of cases including cases such as Psychohacker, and it helps to assess (and deflate) what appear to be potent objections against compatibilist and incompatibilist views of freedom. We believe that these advantages of our account contribute to substantiating the view that authenticity is a condition of responsibility. As we explained previously, the rational credentials of an account are strengthened, other considerations remaining equal, to the extent that the account is theoretically or explanatorily illuminative. In this chapter, we advance further considerations to show that our account enjoys these features. We invoke the relational view of authenticity to resolve two distinct but related problems in the philosophy of education: the problem of indoctrination and the problem of educational authenticity. We begin by commenting briefly on various links between the metaphysics of responsibility (and free action) and the philosophy of education.

We distinguish two different sorts of connection between these two domains of inquiry. First, advances in the metaphysics of responsibility may be used to resolve, dissolve, or illuminate various problems in the philosophy of education. (We do not deny that the converse may hold as well.) The relational account of authenticity, for example, enables us to make progress toward solving the problems of indoctrination and authentic education. Second, the two domains *share* common, deep concerns.

For instance, consider some of David Zimmerman's salient comments on preference acquisition (Appendix A, Section A.3). Zimmerman identifies the chief problem for substantive "positive source-historicism" in the free will debate as "*the puzzle of naturalized self-creation in real time:* How do some children manage to develop the capacity to *make up their own minds* about what values to embrace, by virtue of having gone through a process in which they play an increasingly active role in *making their own minds,* a process that begins with their *having virtually no minds at all?*" (David Zimmerman 2003*a,* p. 638). Explaining the problem further, Zimmerman says, "responsibility-grounding autonomous agency develops with the appropriately *continuous and active participation of the emerging person herself.* The positive historicist thus wishes to clarify the difference between patterns of psychological development that a good liberal would praise as 'education' or 'cultivation,' on the one hand, and condemn as 'indoctrination' or 'psychological manipulation,' on the other" (p. 647). He continues

> The difficulty, however, is to . . . [make] room in the developmental picture for a difference between the kinds of early preference-acquisition that eventually lead to autonomous agency and those that block the child from transcending its early and inevitable heteronomy. This is what I have referred to as the difference between liberal education and authoritarian indoctrination." (p. 655)

We concur with Zimmerman that the educational issues of "authoritarian indoctrination" and "liberal education" are of central concern to the metaphysics of responsibility. We simply add that these concerns are also at the heart of the philosophy of education.

Our relational account of authenticity gives us the conceptual wherewithal to address the dual problems of educational authenticity and indoctrination. We commence with the former.

5.2. THE PROBLEM OF EDUCATIONAL AUTHENTICITY

Appeal to the child's or pupil's authenticity is commonplace in major debates in the philosophy of education. Nuances of the disputes, however, reveal that no evident uniform conception of authenticity informs the dialectic. Different educators or theorists, depending upon the projects in which they are engaged, underscore what seem to be divergent conceptions. We begin with examples that both confirm this multiplicity and highlight the centrality of authenticity in discussions of interest. We then tease out what appears to be a common strand that runs through these seemingly differing conceptions: authenticity is exemplified by motivational elements, such as the agent's desires or values, when these elements are "truly the agent's own"

and not foreign or alien. We argue that it is this sort of view of authenticity that is the mainstay of an important controversy—the problem of educational authenticity—in the philosophy of education: if education, as it appears, entails deliberate molding of the child—it requires, for example, intentional instilment of certain motivational elements in the child—but such intentional molding in the absence of the agent's consent is generally incompatible with authenticity, how is an authentic education even possible? We respond to this problem by invoking our relational account of authenticity, outlined in the preceding chapter, which denies that motivational elements are authentic in their own right; they are authentic only relative to ensuring certain ends.

5.2.1. Appeals to Authenticity

The first set of examples clusters around the theme that autonomy is an educational ideal. Robert Dearden claims that "the development of autonomy as an educational aim . . . is the development of a kind of person whose thought and action in important areas of his life are to be explained by reference to his own choices, decisions, reflections, deliberations—in short, his own activity of mind" (Dearden 1972, p. 70).[1] Dearden's view is a variant of the classical analysis of this ideal whose crux is that an autonomous person makes and rationally assesses her own choices. Richard Peters comments that the classical conception harbors three essential dimensions: intentional choosing, authenticity, and rational reflection. He adds that though being a mentally healthy chooser is a standard expected of normal persons, it is not an educational ideal. Rather, Peters says, in education we are "concerned with the ideal of personal autonomy, which is a development of some of the potentialities inherent in the notion of man as a chooser" (Peters 1973/1974, p. 343). Complementing these reflections, Stanley Benn remarks that to "be a chooser is not enough for autonomy, for a competent chooser may still be a slave to convention, choosing by standards he has accepted quite uncritically from his milieu" (Benn 1976, p. 123). In keeping with the classical analysis, both Peters and Benn insist that autonomous agency requires adopting a code of conduct as one's own and subjecting it to critical scrutiny. Here, the operative conception of authenticity seems closely associated with the capacity to choose guiding principles of conduct on the basis of one's own critical deliberations. Autonomous choice, it is proposed, is thus authentic as well as rationally informed.[2]

The literature on progressivism and the deschooling movement houses the second batch of examples. The classical analysis of autonomy pays homage both to authenticity construed as the capacity to select one's own standards of behavior and to the authority of reason. Educational progressivism, in contrast, makes far more radical demands on the child's authenticity—the relevant conception of authenticity now taking on a distinct slant—insofar as such progressivism combines an appeal to the child's authenticity with

general distrust, or even rejection, of all authority, including that of reason. To appreciate this shift in conception, some background is in order.

One can conceive of childhood as a stage or state, a distinction corresponding roughly, as David Archard remarks, to the distinction between viewing children as "becoming" and "being" (Archard 2003*b,* p. 92).[3] On the first conception, childhood is not yet adulthood and derives its character and importance from this fact. Childhood—customarily divided into the two sub-stages of infancy and adolescence—is thus nothing but a preparation for adulthood. Sometimes this "unfinished person" conception is paired with the view that the primary goal of children's education and schooling concerns transmission to children of the wherewithal necessary for the survival and proper functioning of the society—with its characteristic culture, institutions, and way of life—in which they are born. Under the sway of Jean Jacques Rousseau's *Émile* (1762/1979), "progressivists" call upon an interesting view of authenticity both to criticize this widely held, deeply influential stage conception and to promote its rival that conceives of childhood as a free-standing condition. The view of authenticity invoked is intimately affiliated with a presumption of this free-standing conception that the child has its own characteristic ways of feeling, willing, thinking, and seeing that society or social institutions leave untainted. Child-centered education should do everything, progressivists advocate, to respect these modes of perception.

An especially radical variety of progressivism is the "deschooling" movement. Deschoolers see traditional schools as instruments of coercion, deception, and oppression. In their condemnation of what they regard as the hidden paternalism, indoctrination, and social control of compulsory education, they appeal to yet another conception of authenticity: children are not only the best judges of their own actual needs and interests, they are also the best placed choosers of their own curricula, conformity to which is vital to securing their basic interests. More traditional educational theorists, by and large proponents of the stage conception, dismiss progressivists' and deschoolers' pleas of nonintervention as resting on mere romanticism about children's abilities, and they regard radical noninterference as endangering the healthy mental development of the child.[4]

The last class of examples concerns children's rights and paternalism. Few would dispute the view that children have *some* rights, such as rights against mistreatment, and few would take issue with the submission that parental duties, minimally, include provision of the child with the care and resources for its subsistence and development into adulthood. One controversial issue in the arena of children's rights is the demand to extend to children all or a substantial array of, for instance, the liberty rights adults possess. In their zealous defense of children's rights "child liberationists" yet again appeal to a conception of authenticity. A liberty right presupposes that a bearer of this right has the capacity to choose her religious denomination, vocation, political orientation, citizenship, and so forth. In

this connection, authenticity is allied with a capacity for such choice. Child liberationists insist that children and adults are no different with regard to possession of this capacity. Some opponents of equal rights for children challenge this latitudinarian view on the basis of empirical considerations of child growth. Laura Purdy, for example, suggests that flourishing as adults requires commanding various skills and abilities and mustering self-control. However these skills and abilities do not merely materialize over time; they need active development and training. Allowing children unrestricted freedom to do as they choose within the family, at school, or at work carries the real risk of serious harm: "Granting immature children equal rights in the absence of an appropriately supportive environment would be analogous to releasing mental patients from state hospitals without alternative provision for them" (Purdy 1992, p. 217).

Regarding duties owed to children, the stance of radical progressivists notwithstanding, the limited rationality of children precludes children from determining their own vital interests and taking steps to protect them. Archard (2003*b,* p. 100) explains that the parental obligation to care for one's child bestows upon the parent a power or authority in virtue of which parents have the mandate to be the legitimate interpreters of their children's interests and to make choices for the child that the child is not yet competent to make for herself. The minimal duty of care for children entails that it is appropriate for parents to protect their children against untoward consequences of choices children are thought incapable of making. While not necessarily advocating the extreme paternalism of Hobbes—children are in absolute subjection to parents who may "alienate them . . . pawn them for hostages, kill them for rebellion, or sacrifice them for peace" (Hobbes 1650/1994, 23.8)—some educational theorists have argued for *special* rights of parents over their offspring. These rights do not derive from a prior duty of parents to care for their offspring but are, in some manner, affiliated with the procreative relationship and the attendant naturalness of parental authority. Archard comments that these rights may be viewed as an extension of the parent's rights to lead her life as she chooses, free of interference from others, and they include the "rights to bring the child up in the beliefs, values, and way of life that are the parent's own" (Archard 2003*b,* p. 101) To restrict such parental paternalism, anti-paternal educational theorists invoke a conception of authenticity as the capacity for self-determination to argue, much in the same breath as child liberationists, for the view that children are in actual possession of such a capacity that they can exercise to secure what they judge to be in their best interests.[5]

In sum, important disputes or issues in the philosophy of education frequently appeal to the child's authenticity where authenticity is variously (and non-exhaustively) construed as, or connected with, the capacity to make rational choices about what codes of conduct to adopt, the capacity to feel, will, think, and see in particular ways, the capacity to be the

best judge of one's vital interests and needs, the capacity to exercise liberty rights, and the capacity for self-determination.[6]

We suggest that underlying a substantial range of these divergent threads is a common kernel: authenticity is a property of a person's motivational elements or states that are salient in the generation of her actions. Intuitively, these elements are authentic in that they are "truly the child's own" as opposed to being foreign or alien. In Chapter 3, we proposed that the pertinent contrast between authentic and alien is brought out by reflection on varieties of manipulation that undermine agency or moral responsibility. Shrewd coercion and indoctrination are effective means of getting others, including children, to further one's interests. Desires, dispositions, or habits instilled at the opportune, vulnerable time or over a stretch of time may leave the child without the control that responsibility requires in various spheres of her life. The child, for example, may not be able to refrain from a certain religious practice—her relevant actions would not be appropriately sensitive to reasons—because of the way in which the religious "training" took place. There is a sense in which the germane springs of action that constrain the child's pertinent behavior are alien and not the child's own—actions causally issuing from them are not ones for which the child can shoulder responsibility. Alternatively, recall the dwellers of Walden Two. They seem not to be responsible for their behavior because this behavior is the causal upshot of desires, beliefs, values, and the like that are alien to them. In short, in the sorts of case of interest, the suitably manipulated agent's choices are not free because they issue from elements that are inauthentic or not the agent's own.

We can now see that it is this notion of authenticity—having motivational states or elements that are truly one's own—that appears to unify the otherwise seemingly divergent views previously introduced. So, for instance, if one believes that authenticity is associated with autonomously choosing one's own code of conduct, then as autonomous choice presupposes free choice, and choice is relevantly free only if it issues from authentic springs of action, authenticity construed as the capacity to choose guiding principles of behavior presupposes the more fundamental concept of authenticity as a property of one's motivational springs. Similar things are true with the views of authenticity understood as the capacity to exercise liberty rights, the capacity to determine one's vital interests, or the capacity for self-determination. In any event, it is the notion of authenticity of having motivational springs that are truly one's own, as the second and third set of examples that we outlined strongly suggest, which lies at the heart of a pivotal controversy in the philosophy of education that we wish to address.

5.2.2. Two Extreme Responses

The problem of educational authenticity, recall, can be summarized in this way. As education is a process of molding or influencing, it necessarily

involves interferences; it requires instilling in the child, among other things, salient action-producing elements such as desires, deliberative principles, and values. If the acquisition of these elements totally bypasses the child's capacities of reflective control because these capacities are absent or latent during early infancy, it seems that the child is victim to a kind of subversive manipulation. These instilled elements seem to be just as inauthentic as the ones engineered into the denizens of Walden Two. Progressivists and other theorists of education flag the concern that as the requisite, pertinent educational interferences are of the same genre as those that undermine authenticity, these interferences are incompatible with authenticity. Hence, an authentic education is a pipe dream.

Two extreme responses to this problem can be distinguished and set aside. The *noninterference* Rousseauist model, wedded to the state or "complete little person" conception of childhood, treats the child's "innate authenticity" as sacrosanct, thus regarding all, or if this is a *non-sequitur,* most educational interferences as incompatible with authenticity (Darling 1994, pp. 6–31). A major shortcoming of this hands-off perspective is that the conception of childhood on which it draws is highly suspect. It is not credible, as sundry advocates of this model assume that there is any pre-adult stage at which the child is fully formed, needing only minimal external stimuli to flower into what it is destined to become.

At the other polar extreme is the *nihilist model* that denies that there is anything such as authenticity of motivational springs. Nihilists (or hard paternalists, if one wants) insist that parental and institutional interference in education is *inescapable* and *all there is,* the child being "formed exogenously by the influence of others, as a lump of clay is moulded or a blank slate inscribed upon" (Archard 2003*b,* p. 94). Like the noninterference model, the nihilist view seems overly influenced by a mistaken element of an otherwise promising conception of childhood. This is the element that, at various stages in its development, the child is utterly "unfinished," with motivational and other psychological elements wholly malleable, without any native limits.

In addition, if authenticity in the sense of having motivational springs that are truly one's own is a precondition of free choice, the nihilist model is committed to one of two questionable implications. First, advocates of the model might deny that free choice requires authenticity of salient action-generating motivational elements. However, this flies in the face of both compatibilist and libertarian positions on free action. Both compatibilists and libertarians agree on something like an authenticity requirement but differ in that the former but not the latter provides a compatibilist rendering of the requirement.[7] Second, proponents of the nihilist model might deny that our choices or actions are ever free; lack of authenticity of motivational springs, then, is not an embarrassment because we are not free in the first place. However, there is little reason to accept such skepticism about freedom without weighty argument in its support. More

fundamental to our interests, nihilists themselves have reason to reject this second denial because, just like backers of the noninterference model, they *embrace* the position that a primary goal of education is to ensure that our children develop into morally responsible agents. Responsible agency presupposes the capacity for free choice.

5.2.3. A Reconciliatory Forward-Looking Solution

In what follows, the model we propose as a solution to the problem of authenticity is reconciliatory—some interferences in the process of education are not authenticity subversive because these interferences are *required* for authenticity. Our view is predicated on the principle, advanced in Chapter 3, that there is nothing like authenticity *per se;* motivational elements are not authentic in their own right. Rather, we defended a relational view: they are authentic or inauthentic only relative to whether later behavior that issues from these elements is behavior for which the child is morally responsible. In this respect, our model differs manifestly from the noninterference and the nihilist models, the two being united in affirming that whether there is anything like authenticity *per se* is a mark that divides them. Rousseauists acknowledge, whereas nihilists deny, the existence of authenticity *per se*. Neither entertains the possibility of a relational view of authenticity.

Our model also differs from the outlook of other reconciliationists such as Feinberg and Amy Gutmann. Both Feinberg and Gutmann seem to reject plain authenticity in favor of a relational conception that analyzes authenticity in relation to future adult liberty. On this view, a central goal of education is to work toward the child's becoming an adult who is able to exercise autonomous choice. Educational "interferences," however, are deemed necessary to ensure that children develop into autonomous agents. Paternalist intervention in education is thus legitimate because it is a necessary precondition of subsequent autonomous freedom.[8] This relational view of authenticity, though, is most congenial to those of a liberal outlook who regard autonomy as central to their ideal of the good life. As Archard explains, the Feinberg/Gutmann view will not find favor among those of a communitarian bent who think it important that a child acquire certain values or inherit or continue a certain identity. Elaborating, Archard adds that communitarians do not value the autonomous or chosen life. "What matters to them is tradition, cultural inheritance, or a persisting group identity. They see their children first and foremost as the future members of their group, who must, in consequence, inherit its identity. Here we confront a fundamental difference between liberal and non-liberal understandings of the good life, which communicates itself to views on how best to bring up children" (Archard 2003*b,* p. 100).[9]

Our model, in contrast, is neutral between different conceptions of the good life. It is, in fact, congenial to all interested parties. For it seems that

all parties—communitarians, liberals, Rousseauists, nihilists, and others—concur that a fundamental, overarching goal of education is to make certain that our children develop into morally *responsible* agents. In the conceptual framework that we have introduced, one of education's primary goals is to ensure that our children develop into *morally normative agents*—the sort of agent one has to be if one is to be morally responsible for one's actions. Our view, in brief, is that this goal cannot be attained unless various motivational and doxastic elements—salient action-producing components—are "implanted" in the child. These elements, required to ensure later responsibility for actions that issue from them, are authentic in our relational conceptualization of authenticity. We shall say that these elements are authentic relative to responsibility. Our view on authentic education is in this sense *forward looking:* although pertinent motivational elements instilled in the child during the educational process are not authentic *per se,* they can be "authentic-with-an-eye-toward-future-moral responsibility," not so much despite the necessary interferences on the part of the educators as owing to such interferences. Any such view as ours that claims that various interferences are necessary to assure that the child's motivational springs are relationally authentic, however, assumes the burden of explaining how such interferences—how the instilled motivational elements—differ from ones that subvert relational authenticity. We discharged this burden in Chapter 3. It remains, simply, to apply our results to the issue at hand.

We have proposed that a pro-attitude (or cognitive element, such as a belief) or its mode of acquisition is inauthentic if that pro-attitude (or cognitive element) or the way in which it is acquired would subvert the child's being morally responsible for later behavior that owes its causal genesis to the instilled element. Subversion of moral responsibility would occur as a result of either epistemic, control, or other necessary requirements (independent, of course, of agency presuppositions) of moral responsibility being thwarted. It is in this sense that there can be authenticity-with-an-eye-toward-moral responsibility, but nothing like "plain authenticity" or "authenticity *per se*."

It is important to underscore the precise connection between this relational view of authenticity and moral responsibility. Factors directly pertinent to education underline the connection between the two. As we have emphasized, it is widely accepted that whatever the other goals of education, such as critical thinking, autonomy, or well-being, that one wants to promote (Marples 1999), fully fledged *morally responsible personhood* seems indisputable as one of its overarching aims. Even communitarians, many of whom regard liberalistic education as inimical to a valued way of life, do not—indeed cannot—deny that a pivotal goal of education is to turn children into morally responsible agents. To be a moral agent is to be able, among other things, to participate effectively in the social practices called for by moral responsibility. So, for instance, to become a moral agent, the child must believe that her deliberations, actions, or choices have

upshots in the world, and that she is a fair target of things such as moral praise or blame. (Strawson 1962; Fischer and Ravizza 1998, pp. 210–12) We remarked previously that certain forms of inculcation or nurture are congenial to achieving this goal, other forms, such as paternalism, thwart its realization. We suggested that paternalism threatens attainment of this goal, when it does, principally because it blocks or impedes development of the child into a morally normative agent, an agent who is an apt candidate for ascriptions of moral responsibility.

We proposed that possessing authenticity-destructive pro-attitudes (or doxastic factors) is incompatible with moral responsibility for later behavior which issues from them; possession of such attitudes precludes satisfaction of necessary conditions other than agency conditions required for moral responsibility, such as epistemic or control conditions. Having authenticity-demanding pro-attitudes (or cognitive elements) are required to ensure responsibility for later behavior—having them ensures that necessary conditions other than agency conditions of moral responsibility can (later) be satisfied by the agent or her behavior that stems from them. Authenticity-subversive modes of instilling pro-attitudes or doxastic factors are, like authenticity-destructive actional elements, incompatible with moral responsibility for later behavior. We advanced (in Chapter 3, Section 3.3) the following principle (regarding initial scheme authenticity).

> *Principle of Authenticity:* An agent's initial evaluative scheme is responsibility-wise authentic if its pro-attitudinal elements (i) include all those, if any, that are required to ensure that the agent will be morally responsible for its future behavior; (ii) do not include any that will subvert the agent's being responsible for future behavior that issues from these elements; and (iii) have been acquired by means that will not subvert the agent's being responsible for its future behavior.

All the ingredients for a solution to the problem of educational authenticity are now in place. To ensure that the child matures into a morally normative agent, certain pro-attitudes, beliefs, or values must be "instilled" in the child. However, neither these instilled elements nor their mode of acquisition need subvert the child's being morally responsible, at the age when it can be so responsible, for behavior that causally issues from these instilled elements. Instilling pertinent desires or beliefs is authentic if their acquisition does not subvert, in a characteristic way, moral responsibility for later behavior that (at least partly) issues from these elements. The characteristic way is this: the acquisition of these elements subverts moral responsibility by compromising necessary requirements of responsibility, such as epistemic or control ones, with the exception of agency requirements. If not responsibility-subversive, then these elements are, in the terminology introduced, *relationally authentic*. Some instilled elements or their modes of instilment undercut moral responsibility for later behavior by undermining fulfillment

of necessary conditions of responsibility other than the agency condition. Extreme paternalism, hideously depraving conditions, or experiences traumatic to the child may have this effect. If they do (and empirical evidence is required to confirm whether they do), then in these sorts of case, the instilled elements are relationally inauthentic—they are not "truly the child's own"; the causal pathway to their acquisition is deviant.

5.2.4. Objections and Replies

We now address three major objections. First, one might argue that the position we have defended has a central flaw. It seems to "save" the concept of an authentic education by constructing a notion of "relational authenticity" that appears to be at odds with the intuitive core of authenticity as what is "truly the child's own." Extreme paternalism, hideously depraving conditions, or experiences traumatic to the child may produce an agent moved to do things for which he cannot be properly held responsible, but which are so ingrained that we cannot plausibly say that the springs of action are not his own. The morally servile product of paternalism really *is* servile, even if he cannot be blamed for failing to stand up for what he should. The child who is unwisely beaten will himself become moved by images of violence and inaccurate perceptions of being threatened and must lash out to protect himself. Blaming him will be useless, but the paranoia and violence have become "truly his own." That is what he is like, and the paranoia, violence, and so forth define to a disconcerting extent the only life he will have.

In response, suppose Harris, the victim of physical and psychological abuse in childhood, has motivational springs that move him to aggressive behavior, and that these springs are so deeply ingrained in him that they are "part" of him—they "define" him. We need not assume that *every* such pro-attitude, that owes its genesis partly to his brutal upbringing, undermines responsibility for later behavior that stems from the pro-attitude. The relational view of authenticity will then *not* imply that the pertinent subset of these springs are not "truly Harris's own." In addition, we see no incongruity in the following. A set of pro-attitudes may, in some intuitive sense, "define" what a person is like, and these pro-attitudes may have a pronounced influence on the sort of life he will lead, consistently with not being authentic. To motivate this view, consider this template of a hierarchical account of authenticity (that we do not endorse).[10] The account, in rough strokes, can be profiled in this way. Assume that an agent identifies with a first-order desire of his if he has an unopposed second-order desire that this first-order desire move him all the way to action. The hierarchical view prescribes that one's authentic (first-order) desires are the desires with which one identifies. It may well be that an agent does not identify with an array of first-order desires the members of which habitually move him to action. These desires may "define" him or aspects of him; being akratic in a certain sphere of his life, perhaps he is prone to acting from weakness of will in this sphere. Still,

it appears, there is a clear and intuitive sense in which the desires with which he fails to identify are not truly his own. Similarly, assume that a significant cluster of the first-order desires with which the agent fails to identify are not (relationally) authentic. This fact, in itself, gives us little reason to deny that there is an intuitive sense in which these desires may well "define" aspects of him. Part of the complication here (to which we have signaled above and to which we shall briefly return below) is that there are different conceptions of authenticity, some more salient in certain contexts than others. One conception, for instance, may emphasize what "defines" a person; another, perhaps not entirely divorced from the first and of central interest to us, is crucially associated with free and autonomous choice. We believe that, possibly, the objection at issue may derive the semblance of plausibility at the expense of failing to keep distinct these two (or other) conceptions.

We stress, finally, that authenticity is not all or nothing in that a person's pro-attitudes concerning a certain realm of her activities may be relationally authentic while her pro-attitudes concerning another slate of her activities may not be so. The adolescent's pro-attitudes pertaining to various religious matters may not be relationally authentic because of their mode of acquisition. And yet her pro-attitudes regarding her athletic activities to which she is wholeheartedly devoted may well be relationally authentic. Her pro-attitudes influencing her religious behavior may, again, partly and non-trivially, characterize who she is. However, we see no embarrassment in the suggestion that these pro-attitudes are not authentic when the focus, again, is on decision, choice, or action.

Second, it may be objected that our account of authenticity does not help with cases involving choices of occupation and interests. When we imagine a child not being completely malleable, we imagine, among other things, interests and inclinations or what is now often discussed under the rubric of "intrinsic motivation." A child with a strong urge to make things can be steered around toward practicing law, if that is what her parents want, but the motivation is likely to be "extrinsic" and the life-choice less "authentic," at least with respect to what the child naturally takes pleasure in and could be expected to become good at if given free reign.

As an initial comment, the suggestion that pro-attitudes acquired as a result of being steered by one's parents, when these pro-attitudes are not consonant with those that are "natural" or "intrinsic," are less authentic than the natural ones is controversial. That it is so may be brought out by reflecting on the possibility that the child turns out *not* to be good at doing what she is "naturally" inclined to, she derives *more* pleasure from acting in conformity with her parents' tutelage, and if she were to so act, she would live an overall more fulfilling life. It is not clear to us that, even if one was originally inclined to the view that the natural pro-attitudes are relatively more authentic, one would be so inclined having pondered the possibility just outlined.

Further, the objection seems to presuppose something that is crucially in need of defense: pro-attitudes that are in some sense "natural"—the

"intrinsically motivating ones"—are more truly the child's own than are those acquired as a result of "external influences" such as parental guidance or coxing. This assumption should give us ample reason to pause. For one thing, so-called "natural" pro-attitudes are just as surely acquired as a result of external influences—the environment presumably plays a role in their acquisition as do processes that bear on or involve one's genetic constitution. It would not be credible to deny that one's genetic constitution is the outcome of "external" activities. For another thing, suppose it becomes possible to see to it that our progeny develop certain traits as a result of selective genetic engineering. This possibility seems far-fetched, though not perhaps as far-fetched as once believed. Even if far-fetched, entertaining it will prove useful. Some of what would presumably qualify as the child's "intrinsic motivation" would, on the hypothesis in question, be engineered into the child. Suppose, owing to certain contingencies, the parents have a turn of mind concerning the virtues of letting the child cultivate interests in keeping with the child's engineered-in intrinsic motivation, and they steer the child in another direction. We are hard-pressed to believe that the pro-attitudes that influence the child's interests, and which the parents now try to nurture in the child, are less authentic than the "intrinsic" ones.

Third, one might still have qualms about our relational analysis of authenticity, insisting that it merely constructs a term of art, and does not in doing so salvage the concept of authenticity.

In reply, we begin with a cautionary remark. The essential core of any so-called "alternative analysis" of authenticity which we have suggested is *shared* with the one elucidated: authenticity is relational; there is nothing like "plain authenticity."

Next, the worry appears to assume that there is one appropriate or bona fide concept of authenticity. What this single true or legitimate concept of authenticity is, however, eludes us. As we stressed at the outset of our discussion on educational authenticity, the term "authenticity" is variously interpreted. Some theorists may want to pursue the line that authentic springs are "natural"; others may be drawn to the view that authentic springs are inextricably associated with one's identity. But sundry discussions, including a healthy share in the philosophy of education, appear to center on a notion of authenticity that is vitally connected with the choices or decisions that one makes and with the actions one performs. These are choices or actions that implicate the agent in such a way that, in virtue of doing so (assuming other pertinent conditions satisfied), the agent is an apt candidate for moral responsibility. It is this concept of authenticity that is the target of our inquiry.

5.3. THE PROBLEM OF INDOCTRINATION

We agree with Siegel and others that an ideal of education is to ensure that our children develop into critical thinkers: they should be able to assess

beliefs, desires, actions, and other connative and cognitive elements in their psychological repertoire on the basis of appropriate evaluative standards, be disposed to such evaluation, and be motivated by good reasons in belief-formation and action.[11] We concur, as well, with the ideal that our children blossom into autonomous critical thinkers.[12] Pertinent to this ideal, a salient dimension of being self-governing is that the child matures into an agent who is autonomous with respect to the motivational constituents, such as the desire to evaluate reasons, of being a critical thinker.

The so-called "indoctrination objection," however, casts doubt on whether education, aimed at cultivating autonomous critical thinkers, is possible. The nub of the worry is that the young child lacks even minimal capacities for evaluating reasons. Thus, the constituent components of critical thinking must be indoctrinated if the child is to turn into a critical thinker. Indoctrination, among other things, threatens development of the child into a *self-governing* critical thinker. It is this objection that we seek to defuse. We argue, first, for the view that even if the indoctrination objection can be dealt with at the level of beliefs by an account that distinguishes between beliefs instilled in the child at the non-rational stage that are indoctrinative and those that are non-indoctrinative, there may well be non-autonomous "proto-critical thinkers" who lack autonomy with respect to the requisite motivational components. We then ask what must be added to the account to ensure that proto-critical thinkers develop into autonomous ones. We suggest that motivational elements, even if instilled at a stage at which the child has insufficiently developed cognitive capacities, can be "truly the child's own" or autonomous only relationally: the autonomous motivational elements are ones with respect to which the future child is self-governing and, consequently, ones that may causally issue in later behavior for which the child can then shoulder moral responsibility.

5.3.1. The Basic Issues

Our point of departure is Siegel's reasons conception of critical thinking that views critical thinking as fully coextensive with rationality (Siegel 1988, pp. 32–42; 1997, pp. 2–4). Because both critical thinking and rationality concentrate on the relevance of reasons in believing (or judging) and in acting, critical thinking is rationality's "educational cognate." The reasons conception comprises two related, but conceptually distinct, dimensions: the cognitive *reason assessment* dimension and the motivational *critical spirit* dimension. Respectively, the two are characterized in this way:

> (1) the ability to reason well, i.e. to construct and evaluate the various reasons which have been or can be offered in support or criticism of candidate beliefs, judgments, and actions; and (2) the disposition or inclination to be guided by reasons so evaluated, i.e. actually to believe, judge, and act in accordance with the results of such reasoned evaluations. (Siegel 2003, p. 305)

Elaborating, Siegel proposes that a critical thinker has the ability to assess reasons on the basis of epistemic (and logical) criteria. Reasons appropriately move a critical thinker in thought and action. To be appropriately moved by reasons is, first, to appreciate and accept the importance and evidential force of reasons for beliefs and actions. To determine the relevance and warranting strength of reasons, a critical thinker, moreover, needs to recognize and commit himself to epistemic principles or standards conceived of as universal and "objective." Such standards supposedly guarantee the consistency, impartiality, and non-arbitrariness of reasons. Critical thinking, then, involves the acknowledgment of the binding power of universal and objective evaluative principles in light of which reasons are to be assessed.

Critical thinking theorists distinguish between two sorts of principle of reason assessment: general or subject-neutral principles and context-bound or subject-specific ones. There is an important debate between proponents of a "generalist" view and those of a "specifist" view regarding whether reason assessment skills apply across a broad range of contexts and circumstances: to what extent are assessment criteria generalizable? (Bailin & Siegel 2003, pp. 183–86). Here, we simply note that Siegel adopts the generalist view.

Siegel submits that an agent aspiring to be a critical thinker may have the ability to evaluate reasons but may not systematically exercise this ability. Accordingly, to be appropriately moved by reasons is, second, to be disposed to seek good reasons in support or criticism of candidate beliefs and to question the epistemic credentials of these reasons.

Third, to be duly moved by reasons, a person must habitually and actually engage in reason assessment. Good reasons in belief formation and action must motivate and guide the critical thinker. So, in addition to possessing skills to assess reasons, a critical thinker must have a complex of dispositions, attitudes, habits of mind, and character traits, what Siegel calls a "critical spirit." On Siegel's view, possessing the reason assessment ability and having the critical spirit disposition are individually necessary and jointly sufficient for being a critical thinker.

On Siegel's reason's conception, critical thinking enjoys an impressive generality and wide-ranging relevance in educational contexts. Critical thinking is relevant to, and has ramifications for, the ethics and the epistemology of education, and the content as well as modes of education (Siegel 1988, pp. 42–47). Siegel suggests that we regard critical thinking as an educational ideal, perhaps even *the* ideal of education (Siegel 1997, p. 2). *Qua* ideal, critical thinking not only structures our educational enterprise but also sets the goals of our educational efforts. It gives the answer to two central normative questions in the philosophy of education: how should we educate? and what is education for?

Regarding the means of education, critical thinking operates as a regulative ideal. It defines regulative standards of excellence that can be used to

evaluate, and to adjudicate among, rival educational methods and theories, conflicting teaching methods and theories, alternative curricula, and divergent institutional policies and practices.

As for the end of education, Siegel emphasizes that critical thinking is an "identity-constitutive ideal." The development of critical thinking not only involves inculcating certain reasoning abilities but also inculcating a motivational complex that makes up a certain character. The character traits to be fostered are those constitutive of the critical spirit component. Since having these character traits comprises a model of being a certain kind of person, the fostering of critical thinking is committed to nothing less than the development of a human being with a particular "identity." The fundamental aim of education for critical thinking is, therefore, not only to tutor youngsters to think critically but also, and more comprehensively, to *be* critical thinkers. To take critical thinking as a constitutive ideal is to opt for a pervasive educational program of character-formation and identity-constitution.

Autonomy, in roughly the sense of being self-governing, just like being a critically thinking individual, is frequently thought of as an identity-constitutive ideal: educators should strive to ensure that our children develop into autonomous agents. Indeed, Siegel proposes that there is a sense in which critical thinking and autonomy constitute complementary educational ideals. Critical thinking is, correspondingly, not only closely associated with rationality but also with autonomy (Siegel 1988, p. 54).[13] Siegel writes,

> The ideal [of cultivating reason] calls for the fostering of certain skills and abilities, *and* for the fostering of a certain sort of character. It is thus a general ideal of a certain sort of person whom it is the task of education to help create. This aspect of the educational ideal of rationality aligns it with the complementary ideal of *autonomy,* since a rational person will also be an autonomous one, capable of judging for herself the justifiedness of candidate beliefs and the legitimacy of candidate values.[14]

Elaborating the rational conception of autonomy Dearden and Peters endorse (Section 5.2.1) sheds some light on the alleged complementarity of the ideal of being self-governing and the ideal of being a critical thinker. Recall Dearden's claim:

> the development of autonomy as an educational aim . . . is the development of a kind of person whose thought and action in important areas of his life are to be explained by reference to his own choices, decisions, reflections, deliberations—in short, his own activity of mind.[15]

On this classical conception, an autonomous person makes his own choices and subjects them to rational assessment and criticism. We noted

that Peters ventures that three dimensions of this classical conception are choice, authenticity, and rationality. Being a chooser when exercising practical reason implies having open options and not being restricted by physical or mental impediments. Peters denies that aspiring to be a mentally healthy chooser is an educational ideal. "In education," Peters highlights, "we are usually concerned with more than just preserving the capacity for choice; we are also concerned with the ideal of personal autonomy, which is a development of some of the potentialities inherent in the notion of man as a chooser" (Peters 1973/1974, p. 343). Benn reminds us that a competent chooser may not enjoy autonomy as he may be bound to convention, choosing in accordance with standards uncritically accepted (Benn 1976, p. 123). For this reason, autonomy requires fulfillment of two other conditions. In addition to being a chooser, a person must adopt a code of conduct as his own and subject it to critical reflection in light of rational principles. Autonomous choice has to be authentic as well as rationally informed. Because autonomy on the Dearden–Peters view is so intimately connected with rational reflection, assessment, and criticism, this rationalist conception of autonomy seemingly dovetails with Siegel's reasons conception of critical thinking.

A comprehensive theory of autonomy would account for the autonomy not only of our springs of action, decisions, and overt actions, but also our beliefs (Mele 1995, pp. 86–101), feelings, attitudes, and emotions (Mele 1995, pp. 102–11), and our acquisition, evaluation, and revision of values and deliberative principles (Mele 1995, pp. 112–27). For our concerns, we focus on autonomy with respect to the motivational constituents of critical thinking—the critical spirit dimension.[16] Concerning this issue, Siegel's response to the pressing question, "how can a rational moral code of conduct be acquired by non-rational means?" or, analogously, "how can moral autonomy be created heteronomously?" is instructive (Siegel 1988, p. 86). Siegel appeals to Peters' notion of habit:

> Does the development of proper habits allow us to escape the paradox, and inculcate a commitment to rationality without indoctrinating children into that commitment? It does, if it be granted that habits can themselves become criticizable. If we develop in a child the habit of searching for reasons which justify a potential belief before adopting the belief, that habit not only enhances her rationality; it also admits of rational evaluation itself, for the child can (and we hope will) question the reasons which recommend that habit as a worthy one, and assess the force of those reasons herself. The development of rational habits, then, does not require either indoctrination or the forsaking of rationality. (Siegel 1988, pp. 86–87)[17]

Siegel counsels that the properly educated child cultivates the habit of rational evaluation. In addition, he advises that when the child has the

ability to do so, she critically scrutinize the reasons that recommend this habit as worthy. Nothing in principle, Siegel submits, prevents the child from being autonomous with respect to such habits. To be in the habit of rationally evaluating principles, beliefs, reasons for them, and so forth is, among other things, to be *motivated* to evaluate these items. So it appears that Siegel sees no real concern with the autonomy of the agent relative to the motivational constituents of critical thinking. Further, the passage suggests that Siegel would accept the following constraint.

> *The Critical Thinking Constraint:* If an agent is not autonomous with respect to the motivational elements constitutive of being a critical thinker (such as the desire to acquire or assess beliefs on the basis of evidence), the agent fails to live up to the ideal of being a critical thinker.

We argue that if this constraint is not accepted, it is possible to be a proto-critical thinker who is a slave to reason. Such an agent may acquire and possess beliefs, desires, evaluative principles, and other things on the basis of good reasons, may be disposed to do so, and may act on these critically acquired elements of intentional action but will not be autonomous with respect to the relevant cluster of motivational elements, such as the desire to subject beliefs to rational scrutiny. A proto-critically thinking agent fails to exemplify an *ideal* of education. It should be one of education's primary, overarching aims to strive to ensure that our children develop not merely into (non-autonomous) proto-critical thinkers but into *self-governing* critical thinkers or critical thinkers proper.

5.3.2. The Indoctrination Objection and a Reply

We now turn to the indoctrination objection and to Siegel's response to the objection. These help to bring into sharp relief the distinction between proto-critical thinkers who are non-autonomous in the relevant way and critical thinkers who are pertinently self-governing.

There are different views on what must be going on with regard to *X, Y,* and *p* when *X* is getting *Y* to believe that *p* is rightly thought of as *X* is indoctrinating *Y* into that belief (Snook 1972; Spiecker & Straughan 1991). In the literature on philosophy of education views of indoctrination appeal to either *X*'s intention, or *X*'s method, or *p*'s content, or a selection of these factors, as necessary and/or sufficient conditions. Siegel (1991, p. 30) summarizes the three principal analyses thus:

> One view of indoctrination has it that the case is one of indoctrination if X's *aim* or *intention* is of a certain sort: namely, that X intends to or aims at getting Y to believe that *p,* independently of the epistemic status of or evidence for *p.* A second view holds that indoctrination is a matter of *method,* so that our putative case of indoctrination is a genuine one if

> X's method of getting Y to believe that *p* is of a certain sort: namely, one which tends to impart to Y a belief that *p,* independently of the evidence for *p,* and without Y's questioning *p;* a method, that is, which suppresses or discourages Y's critical consideration of the case for *p.* A third view regards indoctrination as a matter of *content,* so that our case is a case of indoctrination if *p* is false or unjustified, independently of X's intentions and methods.

Siegel proposes that the common denominator of these principal contenders is the fact that the belief is inculcated *independently of the evidence for the belief* so that the believer (*Y*) holds the belief in a non-evidential style. Accordingly, if *Y* holds the belief that *p* without having evidence for it, and if the belief that *p* is not responsive to evidence against it, then the belief that *p* is indoctrinated, whatever might be the intention of *X,* the method of belief inculcation *X* uses, or the content of *p*. In stride with his reasons conception of critical thinking, Siegel offers a non-evidential-style-of-belief conception of indoctrination—or, what he calls, the "upshot" account of indoctrination (Siegel 1988, p. 165, n. 8; 1991, p. 31). A believer who has an evidential style of belief is, in this respect, just like a critical thinker who assesses evidence or reasons for his beliefs. Conversely, if a belief is held non-evidentially, it is not open to rational evaluation and critical assessment. In sum, Siegel proposes that indoctrination is belief inculcation that fosters a non-evidential or non-critical style of belief.

Given this analysis of indoctrination, the indoctrination objection is straightforwardly grasped and seems *prima facie* incontrovertible. In early infancy, the child lacks the cognitive capacities for rationally assessing beliefs, reasons, principles, values, and so forth. In the process of turning the child into a critical thinker, various beliefs, such as the belief that holding beliefs reasons corroborate is preferable to holding beliefs not rationally sustainable, must be instilled in the child. However, the instilled beliefs cannot be supported by the child's critical evaluation of the reasons for these beliefs because the child lacks the concept of reason and he lacks the capacity for critically assessing reasons. The transition from the pre-critical thinking stage of infancy to the stage at which the child has the relevant evaluative capacities is, thus, unavoidably indoctrinative.

Siegel's response to this objection distinguishes indoctrination from properly educational belief inculcation to show that indoctrination in child education is not, after all, inevitable. Siegel admits that in the early stage of infancy, beliefs are inculcated without rational justification on the part of the child. However, at this stage belief inculcation can proceed along two importantly different pathways. Along the first, beliefs are inculcated in such a way that the child is subsequently never encouraged to seek supporting evidence for them and his reason assessment capacity is permanently suppressed. Along the second, beliefs are inculcated "with the view that this lack [of justifying reasons] is temporary, and with an eye to imparting to

[the child] at the earliest possible time a belief in the importance of grounding beliefs with reasons and to develop in her the dispositions to challenge, question, and demand reasons and justification for potential beliefs" (Siegel 1988, pp. 82–83). In this second way, because belief inculcation aims at enhancing the child's rationality and aims for the future "redemption by reasons" of beliefs held *sans* rational justification when instilled, such inculcation qualifies as properly educational belief inculcation. This latter mode of belief inculcation is directed toward development of an evidential style of belief in the child. Since the implantation at an early stage of infancy of pertinent beliefs, deliberative principles, and so on helps to develop in the infant an evidential style of belief, such implantation qualifies as properly educational despite the fact that the young child's capacity for rationally evaluating beliefs is not operative at the time. By contrast, the former mode is the mode of indoctrinative belief inculcation. Indoctrination is a process of belief inculcation that permanently blocks the victim's capacity to think for himself and enduringly prevents him from critically assessing the evidence for the inculcated beliefs. This non-evidential style of believing precludes redeemability by reasons of the indoctrinated beliefs. Siegel concludes that "[t]he indoctrination objection fails to challenge successfully the educational ideal of critical thinking" (Siegel 1988, p. 90).

5.3.3. Proto-Critical Thinkers and Rationality

Now consider these cases. In each, the principal agent satisfies Siegel's requirements for being a critical thinker but is not autonomous with respect to various motivational elements constitutive of the critical spirit dimension of critical thinking. In the first, Ratio develops into a proto-critical thinker, in part, by adoption of an evidential style of belief. Morally questionable means, though, are used to instill the beliefs. For example, the belief that reasons are important, and that acting on the basis of reasons is to be preferred to acting impulsively or without considering the consequences of one's actions, are "beaten into" young Ratio, or inculcated via "shock therapy," or implanted by "exploiting the fear of God's eternal damnation." Desires to acquire beliefs on the basis of warranting evidence, desires not to act precipitously, and other pertinent desires (refer to these as "critical desires") are also instilled in these ways.

In one respect, the inculcation is highly successful: Ratio is transformed into a proto-critical thinker who possesses apposite rational habits. However, one might balk at the immoral techniques used to accomplish Ratio's transformation. A strong concern is that, because these techniques are morally suspect, they intuitively seem to compromise proper education into beliefs. Siegel, though, insists that the method of belief (or desire) inculcation is irrelevant to the distinction between belief (and desire) instilment at the infancy stage that is indoctrinative and belief (and desire) instilment at this stage that is properly educational:

> To focus on *how* the transformation is accomplished, however, is to focus on the wrong concern. The important question is not 'How is the transformation accomplished?'—admittedly, it is accomplished by non-rational means in that the child is not rationally persuaded to become rational—but rather 'Does the transformation, however accomplished, enhance the child's rationality and foster an evidential style of belief? (Siegel 1988, p. 87)

There is good reason to believe that Ratio is not autonomous with respect to the *acquisition* of the critical desires. This compromises his autonomy and lends credibility to the view that, at most, he is a proto-critical thinker. For, appealing to John Christman's insights on the autonomy of acquiring or developing motivational elements or attitudes, Ratio would have resisted acquiring the critical desires in the fashion in which they were acquired, had he attended to their process of acquisition under conditions involving minimal rationality, no self-deception, and circumstances that do not inhibit self-reflection, at a time when Ratio acquired the capacity to do these things (Christman 1991). Further, actions that causally issue from the critical desires are, presumably, actions for which Ratio will *not* be morally responsible when Ratio is a morally responsible agent. This is because these desires *undermine* responsibility for actions Ratio will later perform by preventing satisfaction of necessary conditions of responsibility such as the condition of acting freely. Given the mode of instilling the critical desires, Ratio subsequently finds that he cannot refrain from doing what he perceives to be rationally mandatory.

In the second case, Ratio does not acquire the critical desires via means that are morally objectionable. In addition, he satisfies the historical constraints Christman recommends on the acquisition of desires. Still, Ratio is not autonomous with respect to the *possession* of many of the critical desires. Ratio judges that his quest for evidentially supported beliefs excludes him from acceptance into his religious community. Further, he correctly judges that he would be happier and his life would go better for him if the community were to accept him, and that he would be welcomed only if he were to give up his persistent questioning about the rational credentials of the pertinent religious values, principles, or dictates. Ratio concludes that he should shed his desire to search for evidence for these things. If, despite his judging that he would be better off without this desire, he is incapable, during a span of time, of shedding the desire, then he is not, during that span of time, autonomous with respect to its possession.[18] Since he lacks autonomy regarding the continued possession of the critical desire, his autonomy is, once again, compromised. He is, at best, a proto-critical thinker.

In the third case, relevantly just like the second, if Ratio's desire to search for evidence is uncontrollably powerful, Ratio would not be autonomous relative to the desire's *influence* on his behavior as a critical thinker.[19] Ratio would not, for instance, be capable of exerting even indirect control to prevent

the relevant desires from moving him to action. It is in this sort of case that reason would enslave the agent. In one respect, the agent would be an individual who is an exemplar of an agent who has developed an evidential style of belief. He would have the requisite beliefs, motivational states, and rational habits. However, the agent would be deficient in that he would be non-autonomous relative to the influence of core desires.

Agents such as Ratio in the second and third cases are not the sorts of agent into which we would want our children to develop; we would not want them to become "prisoners of critical thinking." We should aim for a community of *autonomous* critical thinkers and not mere proto-critical thinkers. We submit that the second and third cases provide substantial motivation for the *Critical Thinking Constraint*.

Siegel's response to the worry that his dissolution of the indoctrination objection presupposes that rationality and critical thinking are the ultimate values of a worthwhile life (Siegel 1988, p. 167, n. 24), suggests a challenge to the second and third cases. It may be rejoined that these cases assume that one can have good reason to reject the ideal of reason; Ratio judges that it is best for him to refrain from subjecting various religious values and dictates to rational scrutiny. Similarly, we can imagine an agent who judges that it is best for her to give up an evidential style of belief acquisition and possession altogether. Siegel responds, though, that this assumption of renouncing reason is false. This is because rationality is, in an important sense, self-justifying.[20] Siegel remarks,

> The challenger is arguing, in effect, that there is good reason to reject the ideal of reason. Any such argument against reason, if successful, will itself be an instance of the successful application of reason. That is, the reasoned rejection of the ideal is itself an instance of being guided by it. In this sense, the ideal appears to be safe from successful challenge: any successful challenge will have to rely upon it; any challenge which does not cannot succeed. (Siegel 2003, p. 316)

One may, thus conclude that the second and third cases rest on a presumption that is false; hence, the cases cannot be used to motivate the ideal of autonomous critical thinking.

However, we do not agree that the assumption of renouncing reason, on one construal of this assumption, is false because reason is "self-certifying." We should distinguish between reason and the ideal of being a critical thinker—roughly, the ideal of being a person with an evidential style of belief acquisition and possession. The pertinent question that a Ratio-like agent ponders is the following. Which sort of life should he strive for, a life in which beliefs are acquired and held evidentially or a life in which they are not? Suppose the agent at issue—Ratio in our instance—*reasons* to the second option. (How else, after all, could this question be *non-arbitrarily* settled?) This does not, in any way, sustain the view that the *ideal* of being

a critical thinker is self-justifying. Ratio's choice, on the basis of reasons, shows only that reason recommends abandoning the ideal of being a critical thinker. In the second and third cases, it is this *ideal* that is in question.

It is perhaps worth noting that no incoherence infects the idea that there is a significant sense in which reason itself is not self-certifying. David Gauthier (1986) distinguishes between two different conceptions of rationality, straightforward maximization and constrained maximization. The former is, roughly, the view that an action is rational for an agent if none of its alternatives has a higher expected utility for its agent than it has. Constrained maximization is not as easily formulated. Significantly, though, it differs from straightforward maximization in that, in suitably specified Prisoner's Dilemma contexts, it enjoins that agents opt for interest-constraining yet beneficial outcomes that are beyond the reach of straightforward maximizers. The following matrix highlights these points.

Table 1

		Butch
	Confesses	*Remains Silent*
Sundance		
Confesses	*1,1*	*10,0*
Remains Silent	*0,10*	*9,9*

Straightforward maximizers, Butch and Sundance, are well aware of their predicament as each contemplates the matrix reproduced above. The numbers represent benefits (in utilities) so that more is better. As straightforward maximizers, each knows that no matter what the other does, he does best if he confesses. The outcome of mutual confession, however, is not optimal. (An outcome is optimal if and only if there is no outcome in which some person receives a higher payoff and no person receives a lower payoff.) Each prefers mutual silence to mutual confession. This optimal outcome eludes the straightforwardly rational culprits. If they could only curtail pursuit of their own advantage and refuse to confess, each would be better off.

How would constrained maximizers (according to Gauthier) fare if they were in such a predicament? In parametric contexts where one's choices do not affect others' choices, straightforward maximization and constrained maximization are extensionally equivalent—they generate the same results. In strategic contexts where each interacting agent chooses her action partly on the basis of their expectations of others' choices, constrained maximization requires that

> Each person's choice must be a fair optimizing response to the choice he expects the others to make, provided such a response is available to him; otherwise, his choice must be a utility-maximizing response. (Gauthier 1986, p. 157)

A fair optimizing response is,

> One that, given the expected strategies of the others, may be expected to yield an outcome that is nearly fair and optimal—an outcome with utility payoffs close to those of the cooperative outcome [the (9,9) outcome], as determined by minimax relative concession. (Gauthier 1986, p. 157)

It appears that in strategic contexts where you expect your fellow interactors to cooperate in achieving an outcome that is fair and optimal, then provided such an outcome is possible, constrained maximization requires that you do the "cooperative thing." In those strategic contexts where such an outcome is possible, but where you have no expectation of your fellow interactors cooperating to achieve it, as for instance, could be the case were your fellow interactors straightforward maximizers, constrained maximization requires that you do the straightforwardly rational thing. Constrained maximization tries to ensure that those disposed to cooperate are not taken advantage of by potential exploiters. Finally, in strategic contexts where a fair and optimal outcome is not possible, constrained maximization again requires that you do what is straightforwardly rational. Were Butch and Sundance constrained maximizers, Gauthier proposes, they could rationally secure the cooperative outcome in the Prisoner's Dilemma situation that *Table 1* models.

Directly relevant to our concerns, Gauthier argues that it is coherent for a straightforward maximizer to choose between conceptions of rationality, and that, if rational in the sense of being a straightforward maximizer, such a maximizer would abandon this conception of rationality in favor of constrained maximization (Gauthier 1986, pp. 172–74). Whether Gauthier's intriguing argument is, indeed, successful is not in question.[21] What merits emphasis is the intelligibility of the idea that one conception of reason may rationally be abandoned for an alternative. This, in turn, lends plausibility to the view that an ideal of critical thinking may be rationally abandoned for an alternative life style, and that such a rational choice does not, in any obvious fashion, sustain the contention that reason is self-certifying.

To tie some ends together, according to Siegel, belief inculcation in early childhood with an eye toward the enhancement of rationality and future redemption by reasons is properly educational. Siegel offers a forward-looking solution to the problem of what sets proper education apart from indoctrination into beliefs. We appreciate the power of this solution. It differs from backward-looking solutions which trace indoctrinated beliefs to, for example, belief inculcation in the past that bypasses the agent's capacity for

critically inquiring into the evidential support for the beliefs. We applaud Siegel's insight that significant headway can be made to meet the indoctrination objection, at least at the level of beliefs, by noting that various nonrational ways of instilling beliefs in infants contribute toward development of an evidential style of belief acquisition and possession; beliefs instilled in these ways serve to enhance later rationality. Even if the child, with this training, later acquires the habit of rational evaluation, the resulting adolescent, as the second and third cases involving Ratio confirm, may not be autonomous with respect to the motivational constituents of being a critical thinker. Developing and possessing the habit of rational evaluation, then, will not guarantee autonomy of the requisite motivational components. Ensuring that the agent is autonomous relative to these motivational elements requires treatment different from the treatment that Siegel's appeal to habit recommends. Paralleling Siegel's solution to the problem of properly educational belief inculcation, our solution to the problem of ensuring that children develop into autonomous critical thinkers is also forward-looking. Once again, we apply our relational view of authenticity.

5.3.4. Autonomous Critical Thinkers

Motivational constituents of Siegel's critical spirit component are (relationally) authentic if they satisfy our principle of authenticity:

> *Principle of Authenticity:* An agent's initial evaluative scheme is responsibility-wise authentic if its pro-attitudinal elements (i) include all those, if any, that are required to ensure that the agent will be morally responsible for its future behavior; (ii) do not include any that will subvert the agent's being responsible for future behavior that issues from these elements; and (iii) have been acquired by means that will not subvert the agent's being responsible for its future behavior.

To ensure that the child matures into an autonomous critical thinker, the child must mature into a morally normative agent. To do so requires instilling various pro-attitudes and beliefs in the child. Neither instilled elements nor their mode of instilment need subvert the child's being morally responsible, when it can be so responsible, for conduct deriving from these instilled elements. Instilling pertinent desires or beliefs is authentic if their acquisition does not undermine moral responsibility for later behavior that (at least partly) issues from these elements by subverting necessary requirements of responsibility other than agency requirements, such as epistemic or control requirements. These elements are, then, in the terminology introduced, authentic relative to responsibility. However, some instilled elements or their modes of instilment rule out moral responsibility for later behavior by undermining fulfillment of necessary conditions of responsibility other than the agency condition. To repeat, our view is that

there is nothing like authenticity *per se;* motivational elements, such as desires, are not authentic in their own right. Rather, our relational view of authenticity implies that motivational springs of action are authentic or inauthentic only relative to whether later behavior that issues from them is behavior for which the agent exercises a variety of control, assuming that pertinent epistemic requirements are satisfied.

Briefly, let's now revert to conditions pertaining to autonomously acquiring a desire, autonomously possessing a desire during a period of time, and being autonomous relative to the influence of a desire. We start with the suggestion that one's desires are autonomous *simpliciter* or *sans adjective* only if these desires are "truly one's own" or "authentic." Developmental autonomy can be dealt with in a manner our account of authentic springs of action suggests. At a stage in its development when the child has not yet acquired the capacity to assess reasons—at the pre-initial scheme stage—acquiring a desire is autonomous if its acquisition does not subvert moral responsibility for later behavior that (at least partly) issues (typically, in conjunction with other actional elements) from it. At a stage in its life when a child has grown into a competent reasoner—at the post-initial scheme stage—developmental autonomy requires fulfilling certain history-sensitive conditions roughly of the sort Christman advances.[22] Regarding autonomously *possessing* a desire, we propose that an agent is autonomous relative to the possession of a desire throughout a period of time only if that agent is capable of shedding that desire during that period as a result of exercising the control responsibility requires. Suppose a religious leader implants in Youngster (call Youngster at this age "Infant Youngster") an irresistible desire to act in conformity with his dictates. Suppose later, Youngster ("Elder Youngster") reflects on this desire but is unable to exercise the control responsibility requires to rid herself of this desire. We submit that both Infant Youngster and Elder Youngster are not autonomous relative to the possession of the desire during the pertinent spans of time. As for autonomy concerning the *influence* of a desire, if an agent is autonomous relative to the influence of a desire, it is, in some sense, within the agent's power not to act on that desire; the agent has pertinent control over the action (or actions) that causally issues from the desire.[23] Our yardstick of the type of control to be of the right sort the agent must exercise in performing the pertinent action is whether this control is the control that moral responsibility requires. Since her desire to act in conformity with the religious leader's dictates is irresistible, Infant Youngster will not later be able to exercise responsibility-grounding control in performing actions that issue from this desire. Infant Youngster (like Elder Youngster) is not autonomous relative to the influence of this desire.

In sum, we have proposed that a pivotal aim of education is to safeguard the transition of our children into autonomous critical thinkers. Assume that Youngster's upbringing has equipped her with an evidential

style of belief acquisition and possession and that she has acquired the motivational constituents that Siegel recommends are essential for being a critical thinker. Even if Youngster has cultivated rational habits—she has the habit to assess beliefs, values, judgments, and the like on the basis of good reasons—she may not be autonomous relative to these motivational constituents. Youngster may well be a proto-critical thinker who is reason's slave. Complying with the *Critical Thinking Constraint,* we have proposed a forward-looking account of being autonomous in relation to the constituents of the critical spirit dimension. Add to Youngster's psychological profile that she is autonomous relative to the acquisition, possession, and influence of these motivational constituents, Youngster is then a critical thinker *par excellence*.

5.4. AN UNEXAMINED ASSUMPTION

We argued that an authentic education is feasible and that we have the ability to turn our children into autonomous critical thinkers. Appreciating that these things are within reach is facilitated by clarifying the relevant view of authenticity. Analysis reveals that this view is relational: there is no authenticity *per se* but only authenticity relative to ensuring that the child blossoms into an agent who is, for instance, morally responsible for her later behavior.

However, we have assumed, uncritically, that ensuring that our children develop into *morally* responsible agents and into autonomous critical thinkers are two of education's pivotal goals. We have argued that these goals can be met without "implanting" into our children salient action-producing elements such as desires or values that are "alien" or inauthentic, and without indoctrination. We now examine more closely the first of these assumptions. Perhaps moral responsibility in our lives is not as significant as it has generally been made out to be. If this is so, we may have to reconceptualize what the fundamental goals of education should be.

6 Moral Responsibility, Hard Incompatibilism, and Interpersonal Relationships

6.1. INTRODUCTION: ON THE IMPORTANCE OF MORAL RESPONSIBILITY AND THE SIGNIFICANCE OF LOVE

A seemingly powerful rationale for the proposal that one of education's guiding, overarching aims is to turn our children into morally responsible agents is that moral responsibility is of paramount importance in our lives. This rationale resonates with the views of both compatibilists and libertarians. Though they disagree over whether determinism undermines free action or moral responsibility, advocates of either orientation are generally united in the belief that were the world devoid of moral responsibility, our lives would be seriously morally impoverished. For example, Fischer and Ravizza (1998, p. 3), both compatibilists, contend that, if you were to discover the startling fact that the deliberations, choices, and actions of your best friend are all the product of secretive neuronal manipulation on the part of evil neuroscientists, your most basic attitudes toward your friend would change: your friend would no longer appear to be an appropriate object of such attitudes as respect, gratitude, indignation, and resentment. A lack of moral responsibility seems to threaten some of the moral sentiments and morally reactive attitudes, and in so doing, appears to threaten central interpersonal relationships we greatly value. Fischer and Ravizza (1998, p. 4) additionally theorize that almost everyone would find a life devoid of the morally reactive responses cold and alienating.

Kane (1996, pp. 79–101), an incompatibilist, proposes that without the freedom moral responsibility requires, we could not be ultimate initiators of our actions, genuinely creative, or independent sources of activity in the world. Again, we deeply value things such as vigorous creativity and independence. Further, anticipating some of Fischer's concerns, Kane (1996, p. 65) argues that in the absence of the freedom determinism imperils, we would be like the citizens of Walden Two, agents to whom it would not be appropriate to respond with the morally reactive attitudes.

If moral responsibility is vital, largely because being without it is so costly, we may agree that a pivotal goal of education should be to ensure that our children are nurtured into moral persons—agents who are suitable

candidates for at least some of the morally reactive attitudes on the basis of at least some of their conduct. If, however, the significance of moral responsibility in our lives has been overrated, this allegedly elementary goal of education requires reexamination.

The view that moral responsibility is fundamental to our lives has not gone unchallenged. There are at least two approaches to arguing against it. One approach, which Pereboom elegantly expounds, takes issue with the contention that a conception of life without moral responsibility would be "devastating to our sense of meaning and purpose" (Pereboom 2002, p. 477). Pereboom argues that moral responsibility is not, for instance, required for moral reform and education; for achieving what makes our lives happy, fulfilled, and satisfactory; for healthy interpersonal relationships such as friendship and love; and for an acceptable social policy of criminal behavior. The second approach questions the importance of moral responsibility on the basis of what people actually care about in their lives.

In this chapter, we indicate some shortcomings of Pereboom's approach. The discussion draws on various considerations regarding responsibility for affective states.

6.2. PEREBOOM'S APPROACH: HARD INCOMPATIBILISM

Pereboom endorses a position which he calls *hard incompatibilism*. He argues that moral responsibility for an action depends primarily on its actual causal history and not on the availability of alternative possibilities. Further, he defends the view that, independently of agent-causal accounts of free action, both deterministic and indeterministic causal histories are incompatible with responsibility-relevant freedom. Pereboom claims that agent causalism can accommodate free action and moral responsibility, but there is little reason to believe that we are agent causes. He concludes that no one ever acts freely and so no one is ever morally responsible for anything that one does in this world. Interestingly, though, he argues that a life without responsibility-relevant freedom would not be as detrimental as it has often been made out to be, and in certain respects it may even be beneficial (Pereboom 2002, p. 478).

Filling in some details, Pereboom submits that we can describe cases in which it is evident that a manipulated individual, such as Jenny in Psychohacker, is not free and hence, is not morally responsible for her behavior. A causal history involving apt manipulation, a manipulated causal history, undermines freedom and responsibility. In such cases, the relevant action does not issue from sources over which the agent has control. So the agent is not morally responsible for the action. Pereboom appeals to *Principle O* to support his verdict of unfreedom and non-responsibility in scenarios involving responsibility-subversive manipulation:

> *Principle O:* If an agent is morally responsible for her deciding to perform an action, then the production of this decision must be something over which the agent has control, and an agent is not morally responsible for the decision if it is produced by a source over which she has no control. (Pereboom 2001, pp. 4, 43)

A deterministic causal history, Pereboom contends, is not relevantly different from a manipulated one: it, too, undermines moral responsibility. In scenarios involving determinism, once again, the relevant action issues from sources—the distant past and the natural laws—over which the agent lacks any control. Pereboom proposes that no relevant and principled difference can distinguish an action that results from responsibility-undermining manipulation from an action that has a more ordinary deterministic causal history (Pereboom 2002, p. 478). Pereboom, as we have seen (Chapter 3, Section 3.5; Appendix A, Section A.1), uses a combined counterexample and generalization strategy to argue for this view and, hence, to denounce all compatibilist accounts of freedom.

An indeterministic event-causal history, a history not including agent causation and in which various antecedents of an action, such as the agent's having of reasons, nondeterministically cause elements in the action's etiology or the action itself, is not relevantly different from a manipulated one: it, also, undermines responsibility. This is because in scenarios involving indeterminism, just as in those involving determinism, antecedents over which the agent lacks any control produce the relevant action. Again, Pereboom's position is that no relevant and principled difference can distinguish an action that results from responsibility-undermining manipulation from an action that has a more ordinary indeterminsitic causal history.

Only agent causation allows for moral responsibility. Agent causation is coherent, but given evidence from our best scientific theories, it is not credible that we are in fact agent causes. We, therefore, do not have the freedom that moral responsibility requires.

In sum, Pereboom's argument for hard incompatibilism can be streamlined in this way. Alien-deterministic events are events which factors beyond our control causally determine. Truly random events are events not caused by anything at all. Partially random events are events such that factors beyond the agent's control contribute to their production but do not determine them, and there is nothing, such as agent causation, that supplements the contribution of these factors to produce these events (Pereboom 2001, p. 48). If an action is alien deterministic, truly random, or partially random, then it is not free and hence, no one is responsible for it. This is because, given *Principle O,* no one is the appropriate source of such actions. Every action (mental or otherwise) is alien deterministic, truly random, or partially random (assuming that there are no agent-caused actions). Therefore, no action is free.

Pereboom admits that without this species of freedom, the truth of judgments of moral praise- and moral blameworthiness would be undermined. However, he believes that hard incompatibilism leaves intact other sorts of moral appraisal, such as appraisals of moral obligation, right, and wrong ("morally deontic appraisals"). Pereboom also claims that the hard incompatibilist can defend an acceptable position on managing criminal behavior and on moral education and reform. Finally, he argues that hard incompatibilism leaves interpersonal relationships and "life hopes" significantly unaffected.

6.3. HARD INCOMPATIBILISM, REACTIVE ATTITUDES, AND MORALLY DEONTIC JUDGMENTS

While there is much with which we agree in Pereboom's views, we indicate some concerns. A number of Pereboom's proposals, advanced to soften the blow of being without moral responsibility in, for example, a deterministic universe, presuppose that judgments of moral obligation, right, and wrong (or morally deontic judgments) remain intact in such a universe. Ponder, for instance, some of what Pereboom has to say on moral reform and education. Pereboom entertains the suggestion that even if nobody is ever morally responsible for anything if determinism is true, it would still be best sometimes to *hold* people morally responsible. Such a view might be "justified on practical grounds . . . that thinking and acting as if people sometimes deserve blame is often necessary for effectively promoting moral reform and education" (Pereboom 1995, p. 32). This option, Pereboom remarks, would leave the determinist thinking that someone is blameworthy when she also believes him not to be—an instance of theoretical irrationality—and would have her blaming someone when he does not deserve to be blamed—an instance of wrongdoing (Pereboom 1995, pp. 32–33). Pereboom invites us to consider an option:

> There is, however, an alternative practice for promoting moral reform and education which would suffer neither from irrationality nor apparent immorality. Instead of blaming people, the determinist might appeal to the practice of moral admonishment and encouragement. One might, for example, explain to an offender that what he did was wrong, and then encourage him to refrain from performing similar actions in the future. . . . The hard determinist can maintain that by admonishing and encouraging a wrongdoer one might communicate a sense of what is right, and a respect for persons, and that these attitudes can lead to salutary change. . . . Likewise, although one could not justifiably think of one's own wrongful actions as deserving of blame, one could legitimately regard them as wrongful, and thereby admonish oneself, and resolve to refrain from similar actions in the future. (Pereboom 1995, p. 33, note omitted)[1]

Similarly, addressing some of the reactive attitudes, such as forgiveness, which figure centrally in interpersonal relationships, Pereboom claims that to the extent that the reactive attitudes *do* presuppose blameworthiness, determinism seems to imperil them because determinism undermines blameworthiness. Insofar as our interpersonal relationships depend upon or involve the reactive attitudes, these relationships are also endangered if these reactive attitudes are imperiled (Pereboom 1995, p. 40). However, Pereboom says,

> [T]here are certain features of forgiveness that are not threatened by hard determinism, and these features can adequately take the place this attitude usually has in relationships. Suppose your companion has wronged you in similar fashion a number of times, and you find yourself unhappy, angry, and resolved to loosen the ties of your relationship. Subsequently, however, he apologizes to you, which, consistent with hard determinism, signifies his recognition of the wrongness of his behavior, his wish that he had not wronged you, and his genuine commitment to improvement. As a result, you change your mind and decide to continue the relationship. In this case, the feature of forgiveness that is consistent with hard determinism is the willingness to cease to regard past wrongful behavior as a reason to weaken or dissolve one's relationship. In another type of case, you might, independently of the offender's repentance, simply choose to disregard the wrong as a reason to alter the character of your relationship. This attitude is in no sense undermined by hard determinism. (Pereboom 1995, p. 40)

In addition to his claim that forgiveness presupposes that the person being forgiven deserves blame, in this passage Pereboom suggests that forgiveness also presupposes that the person being forgiven has done *wrong,* and the person doing the forgiving is willing to cease to regard the wrong done to him as a reason to weaken or dissolve his relationship with the person to be forgiven. Similarly, Jeffrie Murphy claims that forgiveness "essentially involves an attempt to overcome resentment," and that resentment—and thus forgiveness—is directed toward responsible wrongdoing (Murphy and Hampton 1988, p. 20). Murphy adds that "if forgiveness and resentment are to have an arena, it must be where such wrongdoing remains intact—i.e., neither excused nor justified" (p. 20). Jean Hampton also agrees that forgiving someone presupposes that the action to be forgiven was wrong (Murphy and Hampton 1988, pp. 40, 54–55).

A key concern with these views of Pereboom to accommodate moral reform and education and primary constituents of reactive attitudes, such as forgiveness, in a deterministic world is that no (true) morally deontic judgments survive in such a world. While we cannot delve into the particulars of why determinism undermines the truth of morally deontic judgments,[2] the gist of the problem is the following.

We assume that determinism expunges alternative possibilities and hence, that determinism precludes its being true that one could have decided or done otherwise.[3] Given this assumption, it suffices to establish the incompatibility of determinism and moral obligation by showing that no act can be morally right or wrong, or obligatory for a person unless that person had the freedom to do otherwise.

We begin by confirming that there is a requirement of alternative possibilities for morally wrong actions. The "ought" implies "can" principle says:

> **OMC**: If it is morally obligatory for one to do something, then one can do it (and if it is morally obligatory for one to refrain from doing something, then one can refrain from doing it).

Another highly plausible deontic principle that links moral obligation and moral wrongness is:

> **OW**: It is morally obligatory for one to do something if and only if is morally wrong for one to refrain from doing it.

OMC and **OW** enjoy both intuitive support and theory-based support; the latter as both are theorems within some of our best theories of the concept of moral obligation.[4] As the ensuing argument establishes, these deontic principles, in turn, yield a third principle that there is a requirement of alternative possibilities for overall wrong actions:

1. If it is wrong for one to do *A*, then it is obligatory for one to refrain from doing *A* (from **OW**).
2. If it is obligatory for one to refrain from doing *A*, then one can refrain from doing *A* (from **OMC**).
3. Therefore, if it is wrong for one to do *A*, then one can refrain from doing *A*.

Barring cogent reasons to believe otherwise, if we assume that "ought" implies "can," there is little reason not to assume, too, that 'wrong" (and "right") imply "can." For the freedom- or control-relevant presuppositions of obligatoriness, it would seem, should also be those of wrongness and rightness. If we grant that "wrong" implies "can," we can show that there is a requirement of alternative possibilities for obligatoriness:

1*. If it is obligatory for one to refrain from doing *A*, then it is wrong for one to do *A* (from **OW**).
2*. If it is wrong for one to do *A*, then one can do *A* (from the "wrong" implies "can" analogue of **OMC**).
3*. Therefore, if it is obligatory for one to refrain from doing *A*, then one can do *A*.

There is no similar way to derive the proposition that moral rightness likewise requires alternative possibilities. For even if it is agreed that "right" implies "can," there is no principle like **OW** that will allow us to infer that "right" implies "can refrain." Nevertheless, it is very plausible that "right" does imply "can refrain." Otherwise, inasmuch as obligatoriness and wrongness do require alternative possibilities, we are in danger of being encumbered with the dubious view that it is morally right for one to do whatever heinous acts one cannot avoid doing.

We can conclude that because there is a requirement of alternative possibilities for actions to be morally right, or wrong, or obligatory, and determinism expunges such possibilities, determinism is incompatible with actions being morally right, wrong, or obligatory. Call any act that instantiates one or more of the primary morally deontic properties of moral rightness, wrongness, and obligatoriness, a "morally deontic act," dub the set of morally deontic acts "deontic morality," and call each proposition that "corresponds" to each of the members of the set, "a morally deontic act proposition."[5] We can now say that determinism is incompatible with deontic morality: if determinism is true, no morally deontic act proposition is true.

Reverting to Pereboom's view on moral reform and education, the alternative he proposes to fostering moral reform by acting as if people deserve praise or blame in a deterministic world *does,* contrary to what Pereboom avows, suffer from irrationality. If determinism is true, nothing is morally right, wrong, or obligatory. Hence, the determinist cannot explain to the offender that what he did was legitimately *wrong,* and then discourage him to perform similar actions in the future. Nor can the determinist, for that matter, communicate to others a sense of what is *right* or *obligatory,* if determinism is true. In sum, if the practice of moral admonishment and encouragement presupposes that actions can be morally right, wrong, or obligatory, a determinist cannot rightly engage in such a practice.

Analogously, if attitudes, feelings, or states such as forgiveness, indignation, resentment, and courageousness presuppose correct morally deontic judgments—judgments to the effect that certain things are morally right, wrong, or obligatory—then as determinism undermines the truth of morally deontic judgments, it also undermines the grounds for such attitudes, feelings, or states. More generally, on the strategy that Pereboom favors, one may seek to salvage remaining constituents or aspects of various things we deem valuable, such as interpersonal relationships (that presuppose the having of certain moral sentiments), moral reform, and various life hopes, in worlds alleged to be bereft of moral responsibility. However if a precondition of these remaining constituents themselves is that the worlds at issue accommodate deontic morality, these constituents will loose their foothold in these worlds if these worlds are deterministic.

Whether various sorts of nondeterministic worlds can accommodate deontic morality is controversial. Should such worlds not be hospitable to deontic morality, then again the strategy Pereboom pursues to mitigate

the alleged detrimental consequences of "living without free will" is somewhat imperiled.

A notable view of Pereboom's, in support of the position that living without moral responsibility is not as bad as it has generally been thought, is that determinism, and more broadly, hard incompatibilism, leaves intact interpersonal relationships including relations of love. In the remainder of this chapter, we cast doubt on this view of Pereboom's. We direct attention, initially, to a spectrum of emotions or attitudes that interpersonal relationships involve and inquire further into whether hard incompatibilism undermines these affections. We rejoin the crucial issue of whether love survives hard incompatibilism in Chapter 9.

6.4. HARD INCOMPATIBILISM, REACTIVE ATTITUDES, AND INTERPERSONAL RELATIONSHIPS

Interpersonal relationships that we deeply value implicate various reactive attitudes. Peter Strawson (1962/1982) proposes that some of the attitudes most important for these relationships are indignation, moral resentment, guilt, forgiveness, gratitude, and mature love. One may attempt to sustain the view that hard incompatibilism threatens interpersonal relations, if one indeed believes that it does, by endeavoring to show that hard incompatibilism undermines the reactive attitudes that are constitutive of, or integral to, these relationships. For example, as Pereboom emphasizes, forgiveness presupposes blameworthiness—when we forgive, the person who is forgiven seeks forgiveness. Owing to hard incompatibilism's subverting blameworthiness (or so, we are assuming), hard incompatibilism subverts forgiveness as well. Thus, a defense of the view that hard incompatibilism leaves interpersonal relations intact may proceed by showing either that some of the reactive attitudes that it has been thought are of vital import to interpersonal relationships are not of such import, or that hard incompatibilism leaves unscathed reactive attitudes or aspects of them that *are* centrally significant to interpersonal relationships. This is the two-pronged strategy that Pereboom adopts.

Pereboom (2001, pp. 207–13), for instance, argues that indignation and moral anger are not obviously required for good interpersonal relationships. We shall, generally, leave Pereboom's attempt to exploit the first prong—showing that reactive attitudes seemingly important for interpersonal relationships, all things considered, do more harm than good—aside. We restrict attention, instead, to Pereboom's case for the view that hard incompatibilism does not undermine the reactive attitudes which he agrees are fundamental to good interpersonal relations.

Regarding this second prong, one of Pereboom's primary defensive maneuvers may be summed up in this way: (i) Certain reactive attitudes and moral emotions play important roles in initiating or maintaining various

interpersonal relations. Good friends, for example, customarily forgive and forgiveness, itself, appears to involve an attempt to overcome resentment.[6] Justin Oakley argues that some emotions are morally significant because they constitute, in various ways, human relationships of love:

> There are several ways in which emotions may be construed as constituting relationships of love and friendship. To begin with, the emotions we both feel towards each other in a sense determine the form our relationship takes. That is, our love or friendship for each other is embodied in our caring about promoting each other's welfare, our feeling sympathetic towards each other in regard to our respective problems, and our feeling angry and indignant at injustices suffered by the other, to name only several. Further, emotions may be thought of as constituting relationships of friendship and love in as far as our mutual affection unifies and bestows a certain significance on our joint activities. We see films and go on walks together out of love and friendship, and many such activities, which might otherwise seem separate and isolated, come to be seen as a complex whole in which our love and friendship are manifested. (Oakley 1992, p. 58)

(ii) Hard incompatibilism undermines constituents of some of the attitudes and emotions of importance to interpersonal relationships. In particular, as we noted in the prior section, a number of these attitudes or emotions presuppose moral praise- and moral blameworthiness which, in turn, conflict with hard incompatibilism. (iii) However, these imperiled emotions or attitudes have either other constituents or, if not, "analogs" that hard incompatibilism does *not* imperil. (iv) These unblemished constituents or these analogs can play the principal, germane roles in interpersonal relationships that the original emotions or attitudes do. (v) So hard incompatibilism leaves secure the pertinent interpersonal relationships.

Further examples should illuminate this interesting defense. Pereboom (2001, p. 201) claims that gratitude may well require the supposition that the person to whom one is grateful is morally responsible for an other-regarding act. Therefore, it may be thought that hard incompatibilism undermines gratitude. However, Pereboom says,

> [C]ertain aspects of this attitude would be left untouched, aspects that can play the role gratitude commonly has in interpersonal relationships. First, gratitude includes an element of thankfulness toward those who have benefited us. Sometimes, being thankful involves the belief that the object of one's attitude is praiseworthy for some action. But one can also be thankful to a pet or a small child for some favor, even if one does not believe that he is morally responsible. . . . In general, if one believed hard incompatibilism, one's thankfulness might lack features that it would have if one did not, but nevertheless, this aspect of gratitude can

> survive. . . . Gratitude involves an aspect of joy upon being benefited by another. But no feature of the hard incompatibilist position conflicts with one's being joyful and expressing joy when people are especially considerate, generous, or courageous in one's behalf. Such expressions of joy can produce the sense of mutual well-being and respect frequently brought about by gratitude. Moreover, when one expresses joy for what another person has done, one can do so with the intention of developing a human relationship. (Pereboom 2001, pp. 211–12)

Remorse and guilt, insofar as they presuppose blameworthiness, also seem to be endangered by hard incompatibilism. It may be ventured that if someone were deprived of remorse and guilt, she would be incapable of mending any relationships with people whom she has wronged; and having done wrong, she would lack any motivation to restore her own moral integrity and thus to develop morally. However, Pereboom recommends that even if one believed in hard incompatibilism, one may feel profound sorrow and regret on being the instrument of wrongdoing despite believing that one was not in any way blameworthy. Sorrow and regret, Pereboom proposes, can play the pertinent roles that remorse and guilt typically do in interpersonal relationships. For example, sorrow and regret may generate a repentant attitude and thus induce the agent not to perform her immoral action again; they may motivate the agent to make amends by seeking to alleviate the suffering caused to others; and they may help to heal the relationship by impelling the agent to express misgiving about her untoward behavior (Pereboom 2001, pp. 205–06). So although gratitude and guilt "would likely be theoretically irrational for a hard incompatibilist," these attitudes "have analogs that could play the same role they typically have" (Pereboom 2001, p. 206).

6.4.1. Responsibility for Attitudes and Emotions

To assess these views of Pereboom, we first distinguish two interpretations of the thesis that hard incompatibilism undermines some reactive attitude or emotion, such as guilt, because this attitude or emotion presupposes moral blame- or moral praiseworthiness, responsibility appraisals that themselves fall victim to hard incompatibilism. Assume that hard incompatibilism is true and call an agent who believes that this demanding position is true a "hard incompatibilist agent." On the first (strong) interpretation of the thesis that hard incompatibilism undermines guilt by virtue of undermining blameworthiness—something that guilt presupposes—the hard incompatibilist agent does not feel anything like guilt, in relevant circumstances, because (as a first stab) she rejects the claim that she is blameworthy and she realizes that guilt presupposes blameworthiness. On the second (weak) interpretation, in relevant circumstances, hard incompatibilist agents *do* have pertinent emotions. Unlike their hard counterparts, they feel something like guilt but the emotion

or attitude is not really guilt because there is a requirement of blameworthiness for guilt. Some may demur, insisting that the agent *would* experience guilt; it is simply that the guilt would be misplaced or irrational. We do not need to settle this issue of whether, on the weak interpretation, what the hard incompatibilist agent feels passes for guilt proper. We shall circumvent it by introducing a term of art. Imagine a situation in which a person appropriately feels guilt. Perhaps the person intentionally does something on the basis of the belief that she is doing moral wrong and that she has no legitimate excuse; she feels guilt upon having done what she takes to be wrong. Now imagine a situation just like this one save that hard incompatibilism is true. Assume that the agent still feels some emotion that would normally qualify as guilt but which some may insist is not really guilt. They would say that it is a "shadow" of guilt because the agent is not blameworthy. Refer to the emotion that she feels as "guilt*." We leave it open whether an instance of guilt* just is an instance of guilt; if we believe that the two are identical, we will say that the guilt is "misplaced," "not well-founded," or "irrational." On the weak interpretation of the thesis that hard incompatibilism undermines guilt (or, more generally, some emotion, E), in relevant circumstances the hard incompatibilist agent feels guilt* (or, more generally, E*) but the emotion that she experiences is irrational, misplaced, or ill founded. We refine the distinction between the strong and weak reading of the thesis in question below. Prior to doing so, we need to say something about responsibility for emotions, feelings, and attitudes.

A full account of free emotions and of moral responsibility for emotions demands an inquiry into the nature of emotions. For our purposes, it suffices to record that if we are morally responsible for some of our actions, then it is highly credible that we are morally responsible for at least some of our emotion-tokens, particular instances of joy or gratitude, for example; we are also morally responsible for at least some of our feeling-tokens, tokens that are affective but that do not qualify as emotions; and we are morally responsible, as well, for, minimally, some of our attitude-tokens, such as tokens of our taking pleasure in various states of affairs. Commenting on the responsibility-relevant freedom or control that we enjoy over some of our feeling states, Mele writes,

> That we have some control over what we feel and over the intensity of our emotions and other feelings is clear. We stem a discomforting flow of sympathy for a character in a film by reminding ourselves that he is *only* a character. . . . The woman who regards her anger at her child as destructive may dissolve or attenuate it by forcing herself to focus her attention on a cherished moment with the child. The timid employee who believes that he can muster the courage to demand a raise only if he becomes angry at his boss may deliberately make himself angry by vividly representing the injustices that he has suffered at the office. . . . These are instances of what I call *internal* control. . . . Many emotions

> and feelings are subject to external control as well—control through one's overt behavior. Jill knows that if, for some reason, she wants to be angry, a phone call to her mother will turn the trick. Jack defeats mild depression by calling his sister. (Mele 1995, p. 106)[7]

We have indirect responsibility-relevant freedom (or control) over something only if we have control over it by virtue of having control over something else. We have direct control over something only if our control over it is not indirect. Similarly, we are indirectly responsible for something only if we are responsible for it via being responsible for something else; we are directly responsible for something only if we are responsible for it but not indirectly so. Responsibility tracks control in that we can be directly responsible for something only if it is in our direct control and if something is in our indirect control we can, at best, be indirectly responsible for it (provided that that thing is not also in our direct control). The control that we have over our emotion tokens or feeling tokens over which we do have control is, presumably, indirect. On pain of avoiding an infinite regress, all indirect control in the end must trace to something over which we have direct control.

Assume that we have direct control only over our decisions. (This is certainly not essential to what is to follow; should one disagree with this assumption, simply supply one's favorite candidate as the candidate for whatever it is over which we have direct control.) If hard incompatibilism is true, though, we have no responsibility-relevant control, direct or indirect, over anything we do and, hence, over any of our decisions. It follows, then, that as we are, at best, only indirectly responsible for, for instance, the consequences of our actions, regardless of the ontological constitution of these consequences, we are not responsible for these consequences. Analogously, it follows that if hard incompatibilism is true, not only are we not responsible for our decisions, we are not responsible for any of our emotion tokens, attitude tokens, or feeling tokens, again, regardless of their constitutional nature, owing to none of these tokens being (indirectly) free.

Assume that our world is deterministic. Hard incompatibilism delivers the verdict that our world is bereft of free action and, thus, bereft of moral praise- and moral blameworthiness. Suppose some agent, upon doing what she takes to be intentional wrong on some occasion, feels guilt* on that occasion. Then we can say, somewhat unfelicitously, that her token of guilt* is not free.

We can now assess the weak interpretation of the thesis that hard incompatibilism undermines a guilt token (or, more generally, some attitude token, feeling token, or emotion token) because this token presupposes that the pertinent agent is morally blameworthy for some germane decision, choice, or action. That Pereboom may favor the weak interpretation over the strong is borne out by passages such as the following:

> How can we deal with our ordinary reactive attitudes—those that are threatened by a belief in determinism—if they are inevitable or extremely

> difficult to alter, yet theoretically irrational and unfair? If we came to the conclusion that hard determinism (or hard incompatibilism) is true, and yet the ordinary reactive attitudes were inevitable, or largely so, it would nevertheless seem inappropriate to maintain the way we regarded those attitudes. . . . Moreover, even apart from considerations of freedom and determinism, most of us have had unfair or irrational attitudes toward others that were difficult or even impossible to eradicate. One is then sad and embarrassed that one has the attitudes in question, avoids indulging or reveling in them, does what one can to rid oneself of them, and one certainly does not justify practical decisions on their basis. This is how it would be best to deal with inescapable resentment and indignation if hard determinism is true, and this way of managing these attitudes does seem to be within our range of capability. (Pereboom 2001, p. 98)

Suppose a hard incompatibilist agent feels a token of guilt* on a particular occasion in her hard incompatibilist world. Either guilt* plays (or can play) the same role that guilt ordinarily does in interpersonal relationships or guilt* does not (or cannot) play this role. Assume, first, that the latter is true. Perhaps guilt* cannot play this role because (i) all guilt* tokens in a hard incompatibilist world are unfree; they are not even indirectly free owing to ultimately deriving from sources over which one has no control. Or, (ii), the hard incompatibilist agent recognizes that all such tokens are irrational in that guilt, if "well-founded," presupposes blameworthiness, but no agent in a hard incompatibilist world is blameworthy for anything. We are imagining, then, for instance, that guilt* will not be adequate for generating a repentant attitude in an agent who has wronged another and, thus, that guilt* will not motivate the agent to refrain from performing such immoral actions again. If guilt*, though, cannot play this role, then it is puzzling why a token of, for instance, sorrow can fulfill the role in interpersonal relationships that guilt ordinarily fulfills because in a hard incompatibilist world each token of sorrow, no differently than each token of guilt*, is unfree, and in a sense, irrational.

To amplify, if I feel sorrow—an alleged analog of guilt—on a particular occasion in a deterministic world because I believe I have wronged you, as hard incompatibilist agents, we would realize that my expression of sorrow is not free and that I am not (even indirectly) morally responsible for this expression. How effective would this token of sorrow be as a vehicle to mending the relationship or as a motivator to restoring my own moral integrity? Hardly at all, we propose, if one bears in mind Pereboom's insistence that a deterministic causal history is not, in principle and relevantly, any different from a manipulated one. My token of sorrow, deterministically caused as it is, might just as well have been implanted in me by nefarious neurosurgeons wanting covertly to control my feeling states. Cognizant of this dubious provenance of all my feeling states, why should the party on the receiving end regard my (unfree) expression of sorrow as "truly mine"

and thus as conducive to healing a wound? If, as it is being assumed, a token of guilt* cannot play the role that guilt ordinarily does because (as noted in (i)) it is not free, then barring contrary reason to believe otherwise, a token of sorrow, too, should not be able to play such a role if it is unfree because of its causal origins.

What of the second reason, (ii), that guilt* cannot play the role that guilt ordinarily does in interpersonal relationships because of its perceived irrationality in a hard incompatibilist world? The concern is that, in such a world, guilt* is irrational because guilt presupposes blameworthiness which is nonexistent in worlds of this sort (or so we are assuming). Unlike guilt, sorrow, after all, does not "presuppose" blameworthiness. Suppose, though, that I am a hard incompatibilist agent. I have betrayed, and in this way, wronged you. Subsequently, I feel sorrow. I realize that the sorrow that I feel is not free; I am aware that the sorrow (or tokens of it) ultimately derives from sources—the distant past plus the laws—over which I have no control. I am further cognizant of the fact that there is no relevant difference between the sorrow that I feel and the sorrow that I would feel if I had been made to feel such sorrow by clandestine direct finagling of my brain. The sorrow that I experience seems just as irrational as the guilt* that I may feel in such a world even though sorrow does not presuppose blameworthiness.

We emphasize that we have *not* argued that sorrow (and for that matter, guilt*) cannot, for example, generate an attitude of repentance in an agent who has wronged another. Rather, our claim is that *if,* in a hard incompatibilist world, guilt* is not up to this role because it is unfree or irrational, then sorrow should not be up to this role either because it, too, is unfree and relevantly irrational. One might want to deny this symmetry but we do not think that Pereboom has sufficiently motivated its denial.

Reconsider, briefly, the analogs for gratitude that Pereboom suggests. In a world that is deterministic Derk benefits me and I am thankful to him. Would this thanking be of value—would it play the customary role in this world that gratitude normally plays? Suppose, further, that I express joy because Derk is especially considerate (as, indeed, he always is!). Would my expression of joy give rise to a sense of mutual well-being and respect? Again, as a (rational) hard incompatibilist, Derk realizes that my expression of joy is a product of factors beyond my control. There is no principled, relevant difference in the way in which the distant past and the laws produce this expression and the way in which this expression would have arisen had it been the product of evil manipulation (or so we are taking for granted). If rational, why should Derk take this unfree expression of joy—an expression that is not "truly mine"—to convey what respect or gratitude customarily convey? Again, bring to mind that under ordinary circumstances, if you discovered that my expression of joy was merely the product of manipulation, your basic attitudes toward me, attitudes adopted partly in virtue of what you took to be a free manifestation of joy, would presumably undergo considerable revision.

To shift to the first horn of the dilemma, assume that guilt*, even if each token of guilt* is unfree and irrational, *can* play the role that guilt usually does in interpersonal relationships. One might propose that suitable analogs of guilt that do not presuppose blameworthiness should, then, also be able to fulfill the role that guilt ordinarily does. In this event, however, invoking analogs to play the role that guilt ordinarily plays would be theoretically enigmatic: the assumption is that the hard incompatibilist agent will feel guilt* on pertinent occasions and that guilt* *itself* can assume the role that guilt does.

Perhaps the concern with the "shadow counterparts" of feeling or emotion states is that our interpersonal relationships, based as they would be on things like guilt*, remorse*, and so forth would be irrational as these emotion or feeling tokens would themselves be irrational. It may be contended that no such irrationality would infect our interpersonal relationships if appropriate *analogs* (sorrow, thankfulness, joy, etc.) that did not presuppose blameworthiness were to take the place of guilt*, remorse*, and so on. However, this latter claim is controversial. The analogs themselves, token expressions of, for example, sorrow and joy, would be unfree and pertinently irrational. If Derk regards the joy that I express in a hard incompatibilist world as originating in factors beyond my control, the joy might just as well have ultimately arisen from the machinations of a naughty cosmic elf. Why, then, in a broad sense of 'rational,' would we be willing to regard such an expression of joy as rational? More generally, if a token of guilt* in a hard incompatibilist world is deemed irrational because well-founded guilt presupposes that the agent is blameworthy for pertinent actions, but that the agent in this world is not in fact blameworthy owing to the causal provenance of her actions, then, similarly, a token of sorrow should be regarded as irrational because it, too, ultimately, derives from sources over which the agent has no control. Again, we find no reason in Pereboom's relevant works to exempt sorrow, joy, or the other analogs from the charge of pertinent irrationality *if* guilt*, remorse*, and so forth fall prey to this charge.

We may reconceptualize the general thrust of our argument so far in the following way. Just as we may distinguish between authentic springs of action, such as desires—desires that are "truly our own"—and inauthentic springs of action, so we may distinguish between authentic sentiments, authentic sorrow, for instance, and inauthentic sentiments. Presumably, the hard incompatibilist would endorse, as a condition of authenticity, that a desire or a sentiment not ultimately originate in sources over which we have no control in order to be "truly our own." The original concern that hard incompatibilism generates for some of our sentiments is that they are irrational (again, in a broad sense of 'irrational') since they presuppose blame- or praiseworthiness that are not possible in a hard incompatibilist world. The hard incompatibilist proposes that the role these threatened sentiments play in interpersonal relationships can be assumed by other sentiments that

are on sure footing in such a world. If, though, there is a legitimate distinction between, for example, authentic and inauthentic sorrow, and all sorrow in a hard incompatibilist world is, according to the hard incompatibilist, inauthentic, then the original concern of hard incompatibilism with respect to moral sentiments, such as guilt, resurfaces at the level of the "replacement sentiments" or analogs.

Let us now evaluate the strong construal of the thesis that hard incompatibilism undermines guilt tokens because guilt presupposes blameworthiness. Recall that on this interpretation, in circumstances in which morally conscientious agents would feel guilt, the hard incompatibilist agent does not feel guilt or, should guilt differ from guilt*, the hard incompatibilist agent does not even feel guilt*. To uncover a pertinent principle on which the strong interpretation is predicated, we introduce the notion of an ultimate originator. Pereboom suggests that if hard incompatibilism is true, we are not the ultimate sources of any of our decisions, choices, or actions because these mental events ultimately originate in sources over which we lack any control. We can say that we are, thus, not the ultimate originators of these things. Now, it might seem that the strong interpretation rests on a principle of this sort:

> SI1: If agent, *S,* is rational, and *S* realizes that (i) *X* is a conceptual requirement of a feeling, attitude, or emotion token, *E,* (ii) *X* (moral blame- or praiseworthiness, for instance) requires the freedom that hard incompatibilism undermines or *S* is not the ultimate originator of *E* or *E*,* and (iii) *S*'s world is a hard incompatibilist world, then if *S* is in circumstances as close as possible to ones in which free action and responsibility are not threatened and in which conscientious moral agents would normally have *E, S* will fail to have *E* or *E**.

The basic idea underlying SI1 is something of this sort. Assume that in a world that accommodates free action and moral responsibility, conscientious moral agents (morally appropriately) feel guilt upon deliberately doing what they believe is wrong and that they lack any excuse for their relevant behavior. Now consider circumstances as close as possible to the ones in which such conscientious moral agents feel guilt but that prevail in a hard incompatibilist world. Rational hard incompatibilists, in these circumstances, would not have any such feeling or emotion. The feeling would extinguish for (partly) the reason that these incompatibilists would believe that guilt presupposes blameworthiness and that their world is shorn of blameworthiness.

But SI1 is false. The mere fact that hard incompatibilism is true, or the mere belief of hard incompatibilist agents in hard incompatibilism, would not ensure that feelings such as guilt or guilt* would extinguish in these agents. Principle SI1 requires supplementation with something to the effect that hard incompatibilists may find it rational to take steps to alter

or eliminate the feelings because these feelings are irrational. Thus, SI1 gives way to the following:

> SI2: If agent, *S,* is rational, and *S* realizes that (i) *X* is a conceptual requirement of a feeling, attitude, or emotion token, *E,* (ii) *X* (moral blame- or praiseworthiness, for instance) requires the freedom that hard incompatibilism undermines or *S* is not the ultimate originator of *E* or *E*,* and (iii) *S*'s world is a hard incompatibilist world, then if *S* is in circumstances as close as possible to ones in which free action and responsibility are not threatened and in which conscientious moral agents would normally have *E, S,* if *S* can, would take steps to eliminate having *E* (or having *E**).

SI2 attempts to capture the following. Consider a world, "Free-World," that is hospitable to free action and moral responsibility. Now consider any situation in which any agent in this world appropriately feels guilt. There is some hard incompatibilist world, "Unfree-World," in which a hard incompatibilist counterpart of each such free-world agent, in an unfree-world situation as close as possible to the free-world situation in which the free-world agent feels guilt, takes steps to extirpate this feeling, provided that this unfree-world counterpart has the ability to do so.

The assumption that even rational hard incompatibilist agents would, perhaps gradually over time, extinguish feelings such as guilt or guilt* is controversial.[8] However, let us assume that their efforts to rid themselves of these feeling states would be successful.

Imagine, now, that you realize that you are the unwitting victim of manipulation; you have been "programmed" to express a feeling token. The token is unfree because of its causal origin; you are not its ultimate originator. If it is rational for you, as a hard incompatibilist, to rid yourself of a feeling token, such as a token of guilt, because guilt presupposes being blameworthy and you cannot be blameworthy for something unless you are its ultimate originator, it seems that it should be equally rational for you to rid yourself of the "engineered-in" feeling, should you be able to do so. In brief, if it is rational for you to eliminate a feeling token because this token presupposes something that is incompatible with hard incompatibilism, it should be rational for you to eliminate a feeling token that is itself unfree if hard incompatibilism is true. Pereboom argues, and for our purposes in this section we are conceding this argument, that in a hard incompatibilist world, there is no principled, relevant distinction between our making a decision as a result of manipulation and our making a decision as a result of, say, this decision's being causally determined. So, as we have previously suggested, by parity of reasoning, there should be no principled, relevant distinction between our expressing a feeling token, such as a token of sorrow or joy, as a result of manipulation or our expressing such a token as a result of its expression's being causally determined. Then if we accept

SI2, we should also accept the verdict that it should be rational for a hard incompatibilist agent to rid herself of tokens of sorrow or joy (because she is not the ultimate originator of these states), if she can rid herself of them. If we assume, further, that rational hard incompatibilists *will* succeed in extinguishing feeling states such as (irrational because unfree) guilt, there should be no bar to assuming that they will also succeed in extinguishing feeling states such as (unfree) sorrow. On the strong interpretation of the pertinent thesis, then, analogs of such feeling states as guilt and remorse will go the way of the feelings states themselves. Consequently, these analogs will not be able to play the same roles in interpersonal relationships that the feelings of which they are analogs ordinarily play.

We previously registered that to secure interpersonal relationships in a hard incompatibilist world, Pereboom argues that some emotional states or reactive attitudes are in fact, all things considered, damaging to interpersonal relationships. We would be better off without such emotions or attitudes. He argues, in addition, that emotional states that *are* vital to interpersonal relationships either have analogs that hard incompatibilism does not debunk or that some of these emotional states themselves are not vulnerable to hard incompatibilism. Having examined the case for survival of analogs, it remains to scrutinize the case for immunity. Concerning the latter, we will circumscribe discussion to Pereboom's proposal that hard incompatibilism does not jeopardize love, which is integral to interpersonal relationships and is, hence, something that we would want to retain. Before we can do this (in Chapter 9, Section 9.2), though, we need to do some preparatory work.

7 On the Significance of Moral Responsibility and Love

7.1. INTRODUCTION: WHAT DO WE CARE ABOUT?

We said in the last chapter that there are at least two different sets of consideration that cast doubt on the view that moral responsibility is pivotal to our lives. The hard incompatibilist marshals one of these, arguing that a life devoid of moral praise- and moral blameworthiness may still be rich and fulfilling. In this chapter, we appeal to the second set of considerations: we devote special attention to what people actually care about in their lives. Essential to this approach is the concept of a non-moral though normative rather than merely a causal variety of responsibility. We explain this concept. We propose that assessments of love, in particular, praiseworthiness from the point of view of love—what we term "commendability"—in contrast to, for example, *moral* praiseworthiness, are especially significant in our lives. We then advance an initial cluster of reasons that tell against love's being immune from hard incompatibilism. We conclude with a defense of the thesis that the importance of *lovable behavior* in our lives is fundamentally tied, in a manner to be explicated, to commendability. This thesis will be instrumental to exposing further reasons against the submission that relationships of love survive intact in a hard incompatibilist world (something we take up in Chapter 9).

7.2. AN ALTERNATIVE APPROACH TO QUESTIONING THE PRIMACY OF MORAL RESPONSIBILITY: OUR CARES AND CONCERNS

An alternative approach to the one Pereboom advances that questions the assumption of the primacy of moral responsibility focuses on the actual behavior of people in everyday life. It appears that the role of moral responsibility in our lives has been overestimated. To develop and defend this view, we introduce the concept of *normative responsibility*. We also appeal to the notions of acting *from duty* and acting *from love,* which we elaborate more fully later.

7.2.1. Normative Agency and Normative Responsibility

Along the lines Bernard Williams suggests, we may distinguish between a narrow conception of morality in which the morally deontic notions of obligation, right, and wrong are primary, and a broader conception in which morality's ambit extends beyond obligation to, roughly, considerations of how one should live.[1] Morality, broadly construed, includes, for example, concerns of love or an ethics of virtue or care. Henceforth, we reserve the use of "morality" or "moral" for the narrow conception. Concerning this conception, we employ "acting from duty" and "acting from moral obligation" interchangeably.

Suppose it is love we deem to be of paramount significance in our lives. An agent can act from love without any thought whatsoever to morally deontic considerations. Such moral factors need play no role at all in the generation of the agent's conduct that is consonant with the requirements of love, such as making significant sacrifices for the welfare of the beloved. Love's requirements may conflict with morally deontic requirements. In cases of such conflict, the agent need not be *morally* praiseworthy for doing what love requires but may, nevertheless, be commendable from the standpoint of love, commendability being an analogue of moral praiseworthiness. Similarly, having acted in light of the belief that she has discharged her moral obligation, though not morally blameworthy, an agent may nevertheless be censurable from the perspective of love, censurability being a genuine variety of responsibility distinct from the moral variety. (Again, bear in mind the identification of the moral with the deontically moral.) In our terminology, commendability (or censurability) from the point of view of love is a non-moral albeit *normative* variety of responsibility.[2] In what follows, "commendability," "censurability" or cognates of the two are terms referring to appraisals of normative responsibility from the point of view of love.

If the duties of love, together with appraisals of commendability and censurability, are prominent in our lives, the focus of education of our progeny should shift from ensuring that children turn into agents who are primarily apposite candidates for moral responsibility to ensuring that they also turn into agents who are suitable candidates for commendability or censurability.

Whatever ideals or duties, such as those of morality or love, are deemed fundamentally important, the corresponding evaluations of responsibility, whether they are of the moral or some other variety, have agency requirements. We take for granted that a person cannot be a fitting candidate for moral praise- or blameworthiness unless she is an apt subject of ascriptions of moral responsibility. Similarly, only suitable subjects can be bearers of commendability or censurability, or for that matter, yet other varieties of non-moral normative responsibility. We may generalize: whatever the variety of normative responsibility, that variety has agency presuppositions. Someone is a *normatively responsible agent* insofar as he is an appropriate

candidate for apt normative attitudes or sentiments on the basis of at least some of his behavior. We concur that a key goal of education is to ensure that our children develop into normative agents. The specific species of normative agency to be given emphasis, at least in the formative, vital years of early education, will be dictated principally by the sort of ideal or standpoint, such as that of love or morality, deemed crucially important in our lives.

To clarify the concept of normative agency, we first expand on the notion of *normative responsibility*. For brevity, we focus primarily on normative blameworthiness.[3] There are different species or varieties of normative blameworthiness. A person can, for example, be morally, love-wise, or prudentially and so normatively blameworthy for intentionally doing or failing to do something or for the consequences of her intentional actions or omissions. Normative blameworthiness is concerned, preeminently, with a certain sort of appraisal of a person and only derivatively with the appraisal of the person's behavior. When a person is normatively blameworthy for an action, the blame in question is inward in that the person is *deserving* of blame, and not "outward." Outward blame includes the outward expression of blame by words, gestures, or actions, and if well substantiated, presupposes blameworthiness.

Normative blameworthiness is closely allied to what a person deeply cares about. Frequently (but not without exception) it is associated with normative standards a person thinks important and, hence, follows in guiding his life and conduct. Construe 'normative standards' liberally. On this expansive interpretation, dictates of custom or tradition, or imperatives deriving from projects or ideals of *central importance* to one's life, count as such standards. Further, for a set of dictates, ideals, or rules to qualify as appropriate normative standards that "ground" normative responsibility, the standards must both guide and constrain behavior; they carry, in the person's life, a sort of authority. A person, who accepts a set of standards as normative, is motivated to act in accordance with those standards, believes that they provide reasons for action, and is disposed to have (appropriate) pro or con feelings or attitudes under various conditions in which the standards are implicated in some fashion. Often (but again not always), when an agent is normatively blameworthy for a course of conduct, the agent does something she takes to be subpar, or below the cared-for normative standards on which she typically relies to arrive at practical judgments about what to do. As an illustration, an agent may do something in violation of *prudential* standards to which she is committed and with which she identifies. She identifies with these standards insofar as she cares more for them than for others such as those of morality or love; it is to these standards she would like her behavior to conform. It is in virtue of the agent's having done something below par that it is frequently fitting, in instances of normative blameworthiness, for the agent to have negative feelings or attitudes (such as regret, or remorse) and for other parties to adopt appropriate negative attitudes toward her; but such feelings on the part of the agent or others are not essential to normative blameworthiness.

It is also worth reemphasizing that the guiding standards with which an agent identifies need not be (deontically) moral. An agent may deliberately evade what she recognizes to be a moral obligation, and intentionally execute some alternative she considers more significant, perhaps because it is the prudentially rational course of action, and because it is prudential standards to which she bears allegiance.[4] Deliberate deviation from such standards may leave the agent susceptible to blame, but the blame will not be moral. We make no presumption that people generally endorse a single set of ideals or standards that guide and constrain behavior across all "domains" of life. One may, with respect to certain concerns, act out of love, but with respect to others, act from moral duty, or from the imperatives of one's religion.

The positive correlate of normative blameworthiness is normative praiseworthiness. Ponder an example involving commendability, a judgment of commendability being a judgment of non-moral normative praiseworthiness. Imagine that a mother visits her sick child in hospital. She sees her child for no other reason than that she loves him and cares for his well-being. The belief, occurrent or dispositional, that it is morally right or morally obligatory for her to visit her child plays no role whatsoever in the etiology of her action or behavior—her visiting her child. Any such moral belief fails to enter into her deliberations (if she deliberates at all) about whether to visit her child; nor in any way does she entertain any moral belief in visiting her child. We submit that the mother is not *morally* deserving of praise for visiting her child. Or suppose that, without hesitation, the mother gives up one of her kidneys to her child who would not otherwise survive. Assume that she acts out of love and not moral duty or any sense of moral concern. Then, again, the loving mother is not morally praiseworthy for giving the kidney. But she is non-morally normatively praiseworthy. She gives up her kidney, roughly, on the basis of the belief that this is what she ought to do. "Ought" in this last sentence does not signal any moral duty or imperative. Rather, it denotes an obligation, or at least some prescriptive element *like* a duty or a deep commitment, associated with acting out of love that is somewhat analogous to what one takes to be one's moral obligation when one acts in light of the belief that one morally ought to do something. The "obligation" here, then, signifies an imperative stemming from the appropriate normative standard from which the mother acts when she gives up her kidney. The standard, in this case of hers, is not a moral one.[5]

Very young children, like our pets, are exempt from responsibility for their conduct partly because they fail to fulfill responsibility's agency presuppositions. One such presupposition is that the candidate be capable of intentional deliberative action. Previously (Chapter 3, Section 3.3), we explained that such action requires some psychological basis for evaluative reasoning. We said that an agent's deliberations which ultimately give rise to some decision or intention, involve the assessment of reasons for or against action by appeal to the agent's evaluative scheme. The constituents of such a scheme include

the normative standards on the basis of which the agent assesses reasons for action. For instance, to be an apt candidate for *moral* responsibility, the normative standards must include a set of moral principles or norms; the agent must be minimally morally competent. Analogously, to be a suitable candidate of commendability and censurability, the agent must have a minimal grasp of the concept of love and its requirements that play a fundamental role in guiding his behavior and in the evaluation of his choices. We explained that the other constituents of an evaluative scheme are the agent's long-term ends or goals that the agent deems worthwhile or valuable; deliberative principles the agent utilizes to arrive at practical judgments about what to do or how to act; and motivation to act on the basis of the normative standards in pursuit of the goals that are elements of his evaluative scheme.

We can now revise the sufficient condition that we previously advanced of being a morally normative agent to arrive at a more general condition that it is not wedded to *moral* responsibility; the general condition appeals to normative responsibility instead. We propose that it is a *sufficient* condition of an individual's being a *normative agent*—an appropriate candidate of *normative responsibility*—at a time *t,* if that individual has at *t* (i) an evaluative scheme with the requisite evaluative elements, moral, prudential, those of love, or yet others—the agent is minimally normatively competent; (ii) deliberative skills and capacities; for example, the agent can assess reasons, values, and so forth by invoking the normative standards that are elements of its evaluative scheme; and (iii) executive capacities—the agent is able to act on at least *some* of its intentions, decisions, or choices. Again, understand condition (ii) to entail that the agent is (at *t*) able to engage in genuine deliberation; her deliberative activities must meet the threshold of rationality below which such activities do not qualify as deliberation.

Depending upon the substantive elements constitutive of a normative agent's evaluative scheme, different varieties of normative agent are possible, moral, prudential, love-responsive, mixed, and so forth. If one primary goal of education is to ensure that our children develop into normative agents, it is pressing to decide what sort of normative agent should be aimed for. We revert to this challenging question in Chapter 9.

7.2.2. The Relative Insignificance of Moral Responsibility

We now address the issue of whether moral responsibility is as significant as it has frequently been thought to be. Our view is that the importance of moral considerations, demands, or concerns in the lives of very many of us has been greatly exaggerated. We are frequently devoted, or more deeply devoted, to other goals, cares, or ideals. Appeal to pertinent examples and reflection on what people actually care about in the relevant sense of "care" motivates this view.

Both Williams and Michael Slote described cases in which a "morally concerned individual might consider a given project to be of greater

importance, for him, than all the harm to other people than that particular project" (Slote 1983, p. 78).[6] Slote develops one of Williams' examples involving a somewhat fictionalized Gauguin:

> We are all to a greater or lesser extent familiar with the fact that Gauguin deserted his family and went off to the South Seas to paint. And although many of us admire Gauguin, not only for what he produced and for his talent, but also for his absolute dedication to (his) art, most of us are also repelled by what he did to his family. . . . I believe that we can persuade ourselves of the wrongness of that desertion and we can do so without losing our sense of admiration for Gauguin's artistic single-mindedness. Single-minded devotion to aesthetic goals or ideals seems to us a virtue in an artist; yet this trait, as we shall see, cannot be understood apart from the tendency to do such things as Gauguin did to his family, and so is not—like daring or indeed like Gauguin's own artistic talent—merely "externally" related to immorality. (Slote 1983, p. 80)

One moral Slote wishes to draw from cases of this sort is that "morality need not totally constrain the personal traits we think of as virtues and there may indeed be such a thing as admirable immorality" (Slote 1983, p. 78). Our interest in the case resides in something different. Spinning the tale as he does, Slote provides convincing grounds for the view that Gauguin's desertion of his family was morally wrong and that Gauguin believed it was so. Nevertheless, deserving emphasis is that Gauguin's passion for art, his zealous devotion to the realization of an "impersonally valuable good," the production of great art that supposedly is of benefit to everyone, took precedence for Gauguin over his concerns for morality (really, immorality) and for his own health or safety. Fictionalized Gauguin is similar in this respect to scores of novelists, sculptors, composers, poets, philosophers, or other scholars. In like manner, some spectacularly (and not so spectacularly) successful business persons, political leaders, professors, or athletes, seeking to accomplish their "professional" goals give less than reasonable weight to their own well-being, and less than normal weight or no weight at all to moral concerns. Their devotion to their projects is not "grounded in," nor does it stem from, any moral obligation or moral concern.

Great artists, novelists, sculptors, and politicians aside, consider one aspect of the relationship among parents and their young. It would be incredulous to believe that the importance to parents of their children's well-being derived in any way from specifically moral obligations to care for their offspring. We care for our children simply because we love them. As Frankfurt explains,

> Moral obligation is not really what counts here. Even if parents are somehow morally obligated to love or to care about their children, it is

> not normally on account of any such obligation that they do love them. Parents are generally not concerned for their children out of duty, but simply out of love; and the love, needless to say, is not a love of duty but a love of the children. To account for the necessities and the authority of parental love, there is no reason to invoke the moral law. (Frankfurt 1994/1999, p. 140).

Similarly, in addition to parental love, there is love among friends and love between spouses, and with love there is the associated care, respect, and trust. For many, there is love for, or devotion to, God. Again, it seems that, at root, moral obligation is not what really matters here. It would false to the facts to suppose that our concern for the well-being of our friends or loved ones somehow derives from specifically moral duties. Here we would do well to remind ourselves of Mill's perceptive remark that ninety-nine hundredths of all our actions are not performed "from" or "out of" moral duty (Mill 1863/1989, p. 23)

The notion of acting from duty is enormously complex. We will return to its analysis later. For immediate interests, we mention one of its central features when it is, first, *moral praise-* or *blameworthiness* or, in short, *moral appraisability,* that is of concern. One is morally praiseworthy for something only if one acts narrowly from moral duty. One acts narrowly from moral duty only if one's act is intentional and the concept of moral duty, understood to encompass moral obligation or moral right, figures pivotally in the representational content of one's intention. So, for instance, a mother acts narrowly from moral duty in giving up one of her kidneys to some child only if she intends to do what she believes morality requires or sanctions. One can act narrowly from moral duty but not purely so as when the mother's act issues partly from an intention to do what she morally ought and to care for the child, where the caring is divorced from any morally deontic association. There are broader conceptions of acting from moral duty. Suppose Theresa is a benevolent person, frequently acting out of kindness. Further, suppose she also has the standing belief that she ought morally to act out of kindness. Now suppose she gives alms to the poor, and in so doing acts out of kindness, but not with any intention to do what she believes morality requires or permits. Her act of giving alms is intentional, but the representational content of the intention Theresa executes when she gives alms is disassociated from any moral concerns. There is a sense in which Theresa acts from moral duty, given her standing belief that she ought morally to act out of kindness. She does not narrowly act from moral duty when she gives alms. Theresa, then, is not morally praiseworthy for giving alms, though other normative assessments, such as that she is virtuous or that her act expresses kindness, are entirely proper.

Still, one might raise the worry that even if we often do not act narrowly *from* moral duty or concerns (and, hence, are often not morally praiseworthy), this alone does not show that we are often not morally *appraisable*.

It needs to be argued, additionally, that we often do not act *despite* moral concerns; for then it would follow that we are often not morally blameworthy—unless we are morally to blame for not "accessing" the relevant concerns or beliefs in the first place.

This worry can, we believe, be met. For first, there are numerous circumstances in life in which we fail to have the relevant moral beliefs—perhaps for the simple reason that we have not thought about what morality requires or forbids in such circumstances. Jan may skip class (indeed, he may form a habit of doing so) without a thought about morality's "informing" his action or cultivation of his habit; a business person may feel that it is "professionally wrong" to divulge company secrets or to take extended coffee breaks but still fail to act, when she deliberatively avoids divulging certain sensitive information or avoids prolonging her breaks, out of or despite *moral* duty or concerns. Second, even in cases in which people have the appropriate sorts of moral belief, they are not morally appraisable for what they do, because these sorts of belief are frequently not "accessed" when they perform the relevant actions. Franz may harbor the dispositional belief that it is wrong for him to be impolite but may, when completely engrossed in his work, snap at interrupters without in any way "accessing" this belief; when discourteous, he acts just as he would have in the absence of having the standing belief; he does not act, even partly, on its basis. Since the counterfactual scenario in which Franz lacks the germane belief but snaps is presumably not one in which he is morally blameworthy for his impoliteness, and it is a scenario relevantly analogous to the actual one in which he is impolite (an appropriate belief is not "accessed" because it is not even possessed), Franz is not morally blameworthy for being impolite. This is perfectly compatible with Franz's being morally to blame, say for failing to "access" the belief that being impolite is wrong or for failing to muster self-control. Third and finally, even in cases involving acting out of love or friendship, or non-moral *concern* for some other individual—such cases are common enough—the agent simply fails to act *despite* moral concern, again for the reason that the agent's behavior in the circumstances is entirely divorced from *any* sort of moral regard or interest.

Suppose we grant that the importance of morality in our lives is limited: very many of our concerns or cares are not in any way derivative from moral duty, and when we act, we often fail to act *from* or *despite* moral duty. Moral concerns—beliefs regarding what is right, wrong, or obligatory, or beliefs that what one is doing is of some moral import—frequently play no role at all in the actual sequence of events that generate our actions. Then the scope of moral responsibility is significantly narrow because moral responsibility requires that we act from or despite moral duty or act "out of" moral concern.[7] Though not morally responsible for many of our everyday actions, we are, it appears, non-morally but normatively responsible for them because we act, for example, from or despite love.[8] Recall our example of the loving mother who is not morally praiseworthy but is

non-morally normatively praiseworthy for giving up her kidney. It is what we take to be important normative concerns in our lives that both restricts the scope of moral responsibility and widens the horizons of non-moral normative responsibility.

7.3. ON THE IMPORTANCE OF LOVE

If moral responsibility is not as significant as it has commonly been thought to be, some other candidate might be of paramount significance in our lives. Commendability and censurability—praise- and blameworthiness, respectively, from love's standpoint—it seems, qualify as apt candidates, their importance being vitally associated with the importance of love. So let us turn to why love is especially valuable.

7.3.1. Love's Value

The notion of *being valuable* is broadly construed as an amalgam of a strict sense of "being valuable" and a derivative sense of this term: something is valuable to, or important for, an agent if it is good (in relation to the agent) in some sense of "good." That is, it is valuable (or good), (i) first, if it is worthy of being valued—it is worthy of being something toward which the agent is favorably disposed. The agent has favorable attitudes, including emotional attitudes, toward it. (This is the strict sense of "valuable.") (ii) Second, it is worthy of being judged good; the agent values it in the (derived) sense of judging, finding, or believing it to be good.[9] (This is the derivative sense of "valuable.")

We propose that love is of value to us. So, too, is lovable behavior which, on a rough and ready approximation, is behavior that is motivated by love. Love and lovable behavior are both (typically) strictly and derivatively valuable to us. Regarding love, in this chapter we summarize some principle reasons for the view that love is derivatively valuable. Later, in Chapter 9, we outline how love contributes to the intrinsic value of lives. The guiding thought to be developed will build on the idea that we take delight in matters of the heart. It is in virtue of our taking intrinsic attitudinal pleasure in the fact that we love or engage in lovable behavior that love is good in a fundamental sense of "good." Love is one object (a pleasure-worthy one) of attitudinal pleasure. So love is worthy of being valued. With respect to lovable behavior, we argue below that, the fact that we (typically) find or judge such behavior to be worthy of being something toward which we are favorably disposed (that is, that we typically judge such behavior to be worthy of being good), is inextricably associated with such behavior's being commendable.

Much of the recent philosophical discussion on love has focused on the analysis of love. On some views, pivotal to the analysis is that there are

reasons for love. On a properties view, the features that constitute reasons for loving a person are a subset of the properties of the person, properties such as *being beautiful, being intelligent,* or *being joyful*. On a relationship view, the reason for loving a person is one's relationship to the person. Rival analyses deny that there are reasons for love. If the fundamental constituents of an analysis of love invoke desires or emotions, and one endorses a thesis that implies that there are no reasons for pertinent desires or relevant emotions, then subscription to a "no-reasons" view will appear attractive. Frankfurt, for example, seems to endorse such a view. He says that love is "essentially a somewhat non-voluntary and complex volitional structure that bears both upon how a person is disposed to act and upon how he is disposed to manage the motivations and interests by which he is moved" (Frankfurt 1999, p. 165).[10] An analysis of love may have a bearing on why love is deemed valuable. However, our concerns require neither that we construct nor defend an analysis. Accordingly, we will not give a definition of love, nor deal with the nature of the different varieties of love. We turn directly to key, fairly non-controversial reasons concerning the value of love. We assume that love is something that is good. Again, we shall propose later that it is good in this fundamental sense: we take delight in the fact that we love. Here, we catalogue principal reasons that have been advanced to support the view that people take love to be, or to believe that, love is good.

Deontic morality is judged to be valuable, among other things, because considerations of moral right, moral wrong, and moral obligation, favorably constrain our behavior. In this respect, love is like morality. Frankfurt manifestly articulates this similarity:

> It is characteristic of our experience of loving that when we love something, there are certain things that we feel we *must* do. Love demands of us that we support and advance the well-being of our beloved, as circumstances make it possible and appropriate for us to do so; and it forbids us to injure our beloved, or to neglect its interests. If we disregard these demands and prohibitions, we feel that we are behaving badly—that we are betraying our love. Now the grip and forcefulness of the requirements that love imposes upon us resemble the forcefulness and grip of moral obligation. In cases of both sorts—those involving love and those involving duty—it seems to us that we are not free simply to do as we please or as we wish; love and duty alike generate in us a sense that we have no choice but to do what they require. In a case of either kind, dereliction on our part both makes us feel that we are somehow at fault and is generally acknowledged to warrant an adverse estimate of our personal character. (Frankfurt 1999, p. 170)

Part of the complex state or condition of love is characteristically trust between the lover and the beloved. Typically, when there is love, there is

supreme or unquestioning trust. Laurence Thomas, for example, claims that one of the distinguishing marks of friendship—and by extension, love—is the bond of mutual trust between friends. This he thinks "is cemented by equal self-disclosure and for that very reason, is a sign of the very special regard which each has for the other" (Thomas 1987, p. 217). Dean Cocking and Jeanette Kennett, take issue with Thomas's proposal that self-disclosure—the confiding of private or intimate information—between friends, cements bonds of mutual trust. But they agree that trust and intimacy are central to friendship. (Cocking and Kennett 1998) Trustworthiness is surely something we value; it is good—we have favorable attitudes toward it—and we judge or believe that it is so.

In a loving relationship, the care of the lover is freely given with nothing expected in return. Frankfurt, for instance, reminds us that lovers identify the interests of their beloveds as their own. He adds that love is concerned with the well-being and flourishing of the beloved object. The good of the beloved is desired for its own sake rather than for the sake of promoting other interests (Frankfurt 1999, pp. 165–66). Similarly, Pereboom affirms that "love of another involves, most fundamentally, wishing well for the other, taking on many of the aims and desires of the other as one's own, and a desire to be together with the other" (Pereboom 2001, p. 202). O. Harvey Green submits that love involves a desire to share an association with the beloved, and that the "basic desire for association motivates and sets parameters for the desire for the good of the one who is loved" (Green 1997, p. 217).[11] Again, such things as the flourishing and well-being of the other and association with the other are things that we deeply value in the derivative sense of 'value'; we judge that these things are good.

In resonance with these views of Frankfurt, Pereboom, and Green; Niko Kolodny tenders that love essentially involves "emotional vulnerability." Kolodny explains that

> To say that A is emotionally vulnerable to B . . . is to say, in part, that A is disposed to have a range of favorable emotions in response to A's beliefs that B . . . has fared or will fare well, and a range of unfavorable emotions in response to A's beliefs that B . . . has fared or will fare poorly. For example, A may feel content when B is well, elated when B meets with unexpected good luck, anxious when it seems that B may come to harm, grief-stricken when B does. (Notice that A is not simply emotionally vulnerable to how B *treats* A, although this is often what is meant by saying that one person is "emotionally vulnerable" to another.) (Kolodny 2003, p. 152)

It goes without saying that love is (derivatively) valuable for other reasons than those we have adumbrated. For example, we value the constancy or resiliency of love. However, the reasons we have mentioned are among the fundamental. We remain neutral on whether an analysis of

love should incorporate the various considerations to which these reasons call attention.

7.3.2. Love Imperiled

We can now appreciate why loving relations may not be as secure as Pereboom suggests that they are in a hard incompatibilist world. Here, it will be useful to divide the discussion into two parts, one concerning hard incompatibilism's impact on *love,* and the other, hard incompatibilism's impact on *lovable behavior,* love being an essential constituent of loving relations and lovable behavior being, characteristically, a part of loving relations.

Beginning with love, toward safeguarding love in a hard incompatibilist world, the primary thrust of Pereboom's endeavors is to convince us that love does not presuppose elements, such as moral praiseworthiness, that conflict with hard incompatibilism. Reconsider, though, Pereboom's affirmation that love of another "involves, most fundamentally, wishing well for the other, taking on many of the aims and desires of the other as one's own, and a desire to be together with the other" (2001, p. 202). Is it true that hard incompatibilism threatens none of these things? Drawing on some of the lessons gleaned from the last chapter, we may ask whether love fundamentally involves, roughly, *freely* wishing well for the other, *freely* adopting apt attitudes or aims, and *freely* desiring to be together with the beloved, where the paradigm of unfreedom is the one Pereboom supplies: adopting or expressing attitudes, having desires, and performing actions (mental or otherwise) are not (directly or indirectly) free if, in the end, they are the product of sources beyond our control. Pereboom is well aware of this potential challenge. He invites us to reflect on "how you would react were you to discover that someone you love was causally determined by a benevolent manipulator to have the love she has for you" (2001, p. 203). Somewhat curiously, having raised this concern, Pereboom initiates an insightful discussion on when the will—making pertinent decisions—intuitively plays a role in generating love for another. He suggests, for example, that a germane decision may be called for to rekindle the intensity of a waning relationship or to do whatever one can to love one's spouse when parents arrange the marriage. Pereboom comments that in such situations,

> we might desire that another person make a decision to love, but it is not clear that we have reason to want the decision to be *freely* willed in the sense required for moral responsibility. A decision to love on the part of another might greatly enhance one's personal life, but it is not at all obvious what value the decision's being free and thus praiseworthy would add. (2001, p. 203)

These remarks do not speak to the *original* concern. We may concede that, first, many of the objects of our love (our children, for instance) are not morally responsible agents, second, rarely, if ever, do we love others

because we intentionally choose to do so, and, third, love engages the will only when certain situations, some of which Pereboom describes, prevail. Still, despite no question whatsoever of loving because of deliberate choice—one is simply captivated by the beloved—there is the issue of whether one loves freely when one loves. Suppose Romeo's love for Juliet is unfree, Romeo being the victim of experimental manipulation. Some of the pertinent desires and affective states typical of love have been engineered into Romeo without Romeo's knowledge of the engineering. Why should Juliet, indeed, why should either party, value this sort of love upon discovery of the exploitation? Surely, it is not *this* variety of love that we cherish.

We may develop this theme by revisiting some reasons why we value love (derivatively). First, we said that part of the complex state of love is trust between the lover and the beloved; trust (or trustworthiness) is something we value. Again it seems that the trust, with its attendant emotional attitudes, must be *free* if it is to be deemed of value. And again, freely trusting each other is something that hard incompatibilism undermines.

Second, we value (at least derivatively) love partly in virtue of the affective intensity or warmth that it requires.[12] As we have seen, Kolodny defends a recent incarnation of this view. In the previous chapter, however, we directed attention to the fact that if hard incompatibilism undermines the freedom of our decisions, it also undermines the freedom of our feeling states. Pereboom's relevant views on freedom imply that there is no principled, relevant distinction between states of emotional vulnerability being causally determined and such states being the product of manipulation. If the emotional vulnerability that love implicates is to be of value to us, it appears that it cannot be vulnerability of the sort that is engineered into us. Why would anyone value this sort of vulnerability?

Third, on some analyses, love is a moral emotion. On others, it is a desire or cluster of desires. On yet others, love is to be analyzed as a relation that has, as an essential element, emotional vulnerability. Whatever the precise nature of love, it is highly credible that love is particular or "nonfungible" in the sense that one would not love another person even if he or she had all the same properties as a person one in fact loves. Simply put, one's beloved is nonsubstitutable. Richard Kraut, for example says, "The non-transferability . . . of love is a defining condition of its being directed toward a unique individual" (Kraut 1986, p. 425). (We return to this nonsubstitutablity theme in Section 8.5.2.)

Now in a hard incompatibilist world, none of our desires is free or authentic and, similarly, none of our emotion or feeling states is free or authentic. This is because, in the eyes of the hard incompatibilist, all of these things ultimately derive from sources over which we have no control. It might be rejoined that even hard incompatibilists should differentiate between, for example, unfree *desires* that are inauthentic and unfree desires that are "truly our own." However, their commitment to the following assumption rules out this option for them:

> *No Difference:* There is no relevant and principled difference between an action whose causal history includes responsibility-undermining manipulation (as in the Ann/Beth scenario) and (i) an action that has a more ordinary deterministic causal history or (ii) an action that has a more ordinary indeterministic causal history. (Chapter 3, Section 3.5, Appendix A, Section A.1, and Chapter 6, Section 6.2)

Actions that arise from inauthentic causal antecedents can be thought of as derivatively inauthentic because their causal antecedents are inauthentic. If actions that derive from responsibility-undermining manipulation of the sort exemplified in the Ann/Beth case, as hard incompatibilists acknowledge, are derivatively inauthentic, and if it is further assumed that more mundane deterministic or indeterministic causal histories are *not* relevantly different from a history involving apt manipulation, there seem to be no grounds to distinguish between the having of desires, the having of which is either deterministically or indeterministically caused, that is authentic, in the relevant sense of 'authentic,' and the having of such desires that is inauthentic.

Assume that love is to be identified with or analyzed in terms of a cluster of desires. Suppose these desires have been engineered into you against your will by the set of new-wave neurosurgeons who worked on Beth. None of these desires is authentic; none is "truly your own." When you love, it is as though it is not *you* who love. Metaphorically speaking, it is just as though you have been "replaced" by someone else who loves. However, if love is nonfungible, such "replacement" cannot preserve love. Why should Juliet still value Romeo's love upon discovering that her lover is, so to speak, substituted by another? In brief, hard incompatibilism seems to fly in the face of the nonfungibility thesis. Similarly, assume that love is a moral emotion or that moral emotions of the pertinent sort are vital constituents of love. None of these emotions of the lover (or the beloved) are truly the lover's (or the beloved's) own if hard incompatibilism rules the day. Once again considerations of the sort just mentioned in connection with the view that love is a desire seem to confirm that love clashes with the nonfungibility thesis in a hard incompatibilist world. We conclude that contrary to Pereboom, there is reason to be pessimistic about whether love survives in such a world.

Even if love itself does not succumb to hard incompatibilism, there are powerful reasons to believe that hard incompatibilism imperils *lovable behavior*. (We use "lovable behavior" and "loving behavior" interchangeably.) We said above that we may tentatively identify such behavior with behavior that is motivated by love. If hard incompatibilism endangers loving behavior, then since such behavior is, typically, a pivotal ingredient of loving relationships, hard incompatibilism threatens such relationships even if it does not threaten love itself.

There are preliminary suspicions about whether lovable behavior remains secure in a hard incompatibilist world. So, for one thing, just as there are moral obligations, so as Frankfurt and others suggest (and as we

have already noted), there are "obligations" or commitments from love's standpoint. We believe that it is intrinsically good when moral obligations are fulfilled and it is, fundamentally, in virtue of this fact that doing moral right for right's sake and shunning what is morally wrong is of intrinsic value and so valuable to us. But, surely, if it is good when moral right is so done, the *presumption* is that some particular moral obligation is *freely* fulfilled—one does not fulfill it as a result of, say, manipulation.

Turning to love, Roger Lamb proposes that as a lover, you are, among other things, obligated *from love's standpoint* to attend to requests of the beloved, help the beloved, be concerned with the welfare of the beloved, and to defend the trust that is partly constitutive of the love (1997, pp. 28–29). We propose that just as it is good when moral right is done, so it is good when love's obligations—obligations or commitments from love's standpoint—are fulfilled. If fulfilling such obligations is good, again the background *presumption* is that these obligations are freely fulfilled. The pertinent sense of 'free' is the sense in which our decisions, for instance, are required to be free if we are to be morally responsible for them. Then the *free* fulfillment of love's obligations is something that hard incompatibilism undermines (or, for present concerns, so we are assuming).

For another thing, since trust (or trustworthiness) and intimacy are fundamental to loving relationships, the behavioral manifestations of these things are part and parcel, typically, of loving relationships. But again, none of these behavioral manifestations is free in a hard incompatibilist world. It is not clear that the unfree behavioral manifestations of trust and intimacy are part of our ideal of what constitutes loving behavior.

So we do think that there are initial, tentative reasons to be somewhat skeptical of the view that hard incompatibilism has no detrimental influence on what is typically a central component—lovable behavior—of loving relationships. In what follows, we develop one line of reasoning to kindle this skepticism. In roughly hewn strokes, it is this: To be lovable behavior, the behavior must exemplify the property of *being commendable* (the property of *being praiseworthy* from the standpoint of love), in contradistinction to, for instance, *being morally praiseworthy*—praiseworthy from the point of view of moral duty. Hard incompatibilism undermines commendability just as it undermines moral praiseworthiness. Thus, hard incompatibilism imperils a crucial component of loving relationships. This line of reasoning, which we take up again in Chapter 9, Section 9.2, requires unearthing a connection between commendability and lovable behavior. It is to this connection that we now turn.

7.3.3. On the Connection Between the Value of Lovable Behavior and Commendability

The burden of this section is to argue for the thesis that *the value of lovable behavior for us is essentially a function of our being commendable for the*

behavior. This thesis is to be understood as implying the following. First, behavior is *lovable behavior* only if its agent is commendable for the behavior. If an agent's actions are in accordance with the requirements of love but the agent is not commendable for those actions, then the prior implicate yields the result that the behavior will not be lovable behavior and, hence, it will be behavior that is devoid of the value we typically associate with loving behavior. Second, it is in virtue of possessing the feature of *being commendable* that lovable behavior is especially valuable.

To understand the view that lovable behavior is behavior for which its agent is commendable, we need to tighten up on our account of what species of behavior we have in mind when we speak of lovable behavior. Refer to intentional behavior that is in accordance with love's requirements but that is behavior (an intentional action, for example) for which its agent is not commendable as "loving* behavior" (or as "a loving* action") or, if one wants, as "ersatz loving behavior." Assessments of love's *requirements* or *prohibitions* are assessments of behavior that are "act focused"; such assessments are first and foremost normative appraisals of the *behavior* and not appraisals or appraisals only derivatively of its agent. Assessments of commendability, in contrast, just like assessments of moral praiseworthiness, are primarily "agent focused"; they are fundamentally normative appraisals of the agent and not, in the first instance, appraisals of the pertinent behavior. To be behavior that is *loving behavior,* the behavior must be expressive of love. (Hence, the initial tentative gloss that lovable behavior is behavior that is motivated by love.) To be expressive of love, its agent must be commendable for the behavior; the behavior must be reflective of the loving attitude of the agent toward the beloved. Thus, loving behavior, as we understand it, is behavior that is in accordance with the requirements of love and for which its agent is commendable.

The stance toward which we are working is this: Given the notion of *being valuable* at issue, if loving behavior is good—it is behavior worthy of our having appropriate favorable attitudes toward it—and we take such behavior to be good, then such behavior is important to us; it is of value to us. There is little reason to believe that, generally, people are favorably disposed toward *loving* behavior;* such behavior is not typically behavior worthy of our having favorable attitudes toward it. People do not, for example, generally, take delight in engaging in loving* behavior. In addition, there is little reason to believe that people typically find loving* behavior to be good. There is, thus, little reason to believe that loving* behavior is good in one fundamental respect in which *loving behavior* is good: we take delight in the fact that we engage in the latter sort of behavior but, generally, we do not take pleasure in engaging in the former sort of behavior. In sum, there is little reason to sustain the view that loving* behavior is good and that people take loving* behavior to be good. It follows that loving* behavior is not (typically) of value to us. It is loving behavior proper, behavior that entails commendability, which is valuable to us. Roughly, it

is the agent's "proper (loving) investment" in a bit of loving behavior that we cherish so deeply.

The thought, that loving* behavior, that is, again, behavior in accordance with the requirements of love but unaccompanied by commendability, is not the sort of behavior we have in mind when we think of loving behavior as valuable, may be developed, in a preliminary fashion, by reflecting on Pereboom's remarks on love that we cited previously: love of another involves, most fundamentally, wishing well for the other, taking on many of the aims and desires of the other as one's own, and a desire to be together with the other. One may wish well for the other because one believes that this is morally or prudentially required of one. Similarly, one can take on many of the aims of the other as one's own, or generate desires to be together with the other, or sustain such desires, because one believes that this is what morality requires. What we would then value in such behavior, if we value it at all, would not be anything like what we value in loving activity. What we find valuable in behavior of this sort, insofar as such behavior is genuinely lovable behavior, is that the relevant agent—the lover, for instance—*is* commendable for the behavior. The behavior expresses the cares or nuances of *love*. To elaborate, we remarked that when one loves another, one is typically concerned for the other. The concern may express itself in sundry ways, many behavioral. Insofar as the concern is a concern of love—insofar as the behavior that expresses the concern is genuinely lovable behavior—what is done to manifest the concern, it seems, causally stems appropriately from love and not, for example, from duty or prudence—the behavior must be behavior for which one is commendable. Adapting an example of Williams, the spouse, saved by the husband who declares that he rescued his wife partly in view of the fact that that is what love required of him, but who failed to act on the basis of the belief that love constrained him to act in the way in which he did and, so, who failed to act "out of" love, would be just as put off as she would have been had her husband informed her that he acted solely from moral duty in saving her. The husband acted in conformity with the requirements of love, but not being commendable for his behavior, we would be hard pressed to regard his behavior as loving.[13]

Cocking and Kennett propose that a close friend—a lover, for instance—is receptive to being directed and interpreted by the other. They explain that when one is directed in the characteristic way, "one's choices are shaped by the other and one's interests and activities become oriented toward those of the friend" (Cocking and Kennett 1998, p. 504). In an example that they develop, on the basis of one's receptivity to being directed by one's friend's interests, one accepts the friend's invitation to the ballet even though one has no interest and will never have any real interest in the ballet. In acting out of love or friendship, one does not go begrudgingly or out of any sense of moral obligation (p. 504). Yet again, though, we would not find anything of value commensurable to what we find of value in loving behavior, if one were to go to the ballet but not be commendable for doing so.

Reflecting on receptivity to being interpreted by the other, Cocking and Kennett advance the following case.

> Consider how we often recognize and highlight aspects of our close friend's character. So, for example, Judy teasingly points out to John how he always likes to be right. John has never noticed this about himself; however, now that Judy has pointed it out to him he recognizes and accepts that this is indeed a feature of his character. Seeing himself through Judy's eyes changes his view of himself. But beyond making salient an existing trait of character, the close friend's interpretation of the character trait or foible can have an impact on how that trait continues to be realized. Within the friendship, John's liking to be right may become a running joke which structures how the friends relate to each other. John continues to insist that he is right; however, his insistences are now for the most part treated lightheartedly and take on a self-consciously ironic tone. And John may be led by Judy's recognition and interpretation of his foibles to more generally take himself less seriously. Thus, John's character and his self-conception are also, in part, drawn, or shaped, by his friend's interpretations of him. (Cocking and Kennett 1998, p. 505)

If Judy were not commendable for bringing the indicated foible to John's attention, we would suspect that she is not acting out of friendship or love. Her behavior, at best, would qualify as ersatz lovable behavior. Analogously, suppose John reacts to Judy's activities in the way in which Cocking and Kennett describe in the passage. Again, if John were not commendable for the pertinent behavior that comprises his reactions, we would have good grounds to believe that he did not act out of love or friendship. His behavior would be devoid of what we find valuable in loving behavior.

What, though, about cases in which one loves seemingly without exhibiting any *overtly behavioral* manifestations of love? *For* what, in such cases, is the agent commendable? Velleman, for instance, brings attention to scenarios that suggest cases of the relevant sort:

> [S]urely, it is easy enough to love someone whom one cannot stand to be with. Think here of Murdoch's reference to a troublemaking relation. This meddlesome aunt, cranky grandfather, smothering parent, or overcompetitive sibling is dearly loved, loved freely and with feeling: one just has no desire for his or her company. . . . In the presence of such everyday examples, the notion that loving someone entails wanting to be with him seems fantastic indeed. (Velleman 1999, p. 353)

Similarly, Velleman suggests,

> I think that one can love a person without having the faintest notion of what that person's interests are, and without having any inclination to

> discover or promote them. One may feel unworthy to serve the beloved's interests, or powerless to serve them, or forbidden from serving them by social circumstances or ethical constraints. One may love a colleague or student in ways that one is not entitled to express in benevolent action. One may love a teacher or mentor without ever presuming to imagine that one might further his interests. There are even loving friendships, I think, in which respect for one's friend rules out any acts of unsolicited benevolence. (Velleman n.d., p. 18)

Cases such as these, though, do not present any substantial difficulty for the thesis at issue. Surely, a person may express loving feelings and may well be commendable for expressing such feelings. Or if, as Velleman believes, "love is essentially an attitude toward the beloved himself but not toward any result at all" (Velleman 1999, p. 354), there is nothing, in principle, to stand in the way of the person's being commendable for the attitude or appropriate constituents of it. Indeed, as Michael Zimmerman forcefully argued, if we do not conflate the scope of moral responsibility—roughly, the things *for which* an agent is morally responsible—with degree of moral responsibility—roughly, the *extent to which* a person is morally responsible—then there is nothing untoward about a case in which the scope of, say moral praiseworthiness, diminishes to naught but in which the degree of such praiseworthiness remains the same as what it is in an otherwise similar case in which the scope of moral praiseworthiness is significant (M. J. Zimmerman 2002) There is no reason to think that commendability differs from moral praiseworthiness in this respect. Thus, a person can be commendable for her loving attitude although she does not in any way overtly manifest this attitude in loving behavior; and she can be commendable for it to the same extent as she would have been had her attitude found expression in loving behavior.

In summary, should we be taken to task to clarify the general line of reasoning to sustain the thesis that what we find valuable in loving behavior is essentially a function of our being commendable for the behavior, we oblige with the following. First we distinguish between behavior that is merely in accord with the requirements of love (ersatz lovable behavior) and genuinely loving behavior. We record the truism that we typically value the latter but not the former. The explanation of why we customarily value the latter is, again, the relatively straightforward one that it is *lovable* behavior that is characteristically valued. We then ask what it is about such behavior in virtue of which it qualifies as lovable behavior proper as opposed to qualifying merely as ersatz lovable behavior. We take our cue from suggestions such as the following. When an agent engages in ersatz lovable behavior, this behavior does not express the cares or concerns of love; the behavior need not causally stem from desires for the good of the other for the other's own sake; or the behavior does not generally express the "investments" of love, such as taking on many of the aims and desires

of the person who is loved as one's own. We propose that underlying these suggestive reflections is the unifying view that lovable behavior (however thoughtful or reckless) is behavior for which its agent is commendable. If an agent is commendable, for example, for an action, then she performs that action at least partly on the basis of the belief that that is what *love* requires that she do. Given that all other conditions of commendability, such as freedom-relevant conditions, are satisfied, the agent will be commendable for this action.

We conclude that there are strong reasons to believe that the thesis under scrutiny that ties the value of lovable behavior to commendability is on sound foundations. After formulating and rebutting objections to this thesis (something we do in the next chapter), we complete the argument for hard incompatibilism's undermining lovable behavior in Chapter 9, Section 9.3.

8 Love, Commendability, and Moral Obligation

8.1. INTRODUCTION: IN DEFENSE OF COMMENDABILITY

In this chapter, we first defend the view that the value of loving behavior to us is essentially a function of our being commendable for that behavior against various objections. We then turn to critical scrutiny of a thesis that we accept and that we have invoked in our previous discussion on love: it is possible for an agent to perform an act that issues from love—the agent can act out of love—without the act's issuing from duty—without, that is, the agent's acting from moral duty or obligation.

8.2. LOVE AND COMMENDABILITY: AN OBJECTION

Love is valuable or important to us and it is so for many reasons. We have focused, in part, on the value for us of loving behavior. We have argued that what is valuable to us in behavior that love requires of us, is inextricably associated with our being commendable—praiseworthy from love's standpoint—for this behavior. If we value behavior that is merely in accordance with the requirements of love but for which we are not commendable, what we value in such behavior is not what we value in behavior that is loving behavior proper.

One may question the thesis that the value of lovable behavior, for instance, the value we find in acts expressive of love's concerns, is essentially a function of being commendable for these acts by drawing our attention to acts that are expressive of desirable traits or virtues. Prompted by compassion to do so, Tania helps someone in distress. The virtue of being compassionate figures centrally in her motivation to act as she does. Tania is positively evaluable *vis-à-vis* performance of her act that causally stems from compassion. It may reasonably be submitted that we value acts of compassion because the agent is so positively evaluable in relation to them; when one acts from compassion, one's act expresses a virtue. It may further be proposed that the value of lovable acts should be conceptualized in an analogous fashion.

The proposal, though, is mistaken. We grant that Tania is positively evaluable. However, we direct attention to the favorable moral appraisal at issue being an aretaic one. We may say that Tania's act has "moral worth" insofar as it expresses the virtue of compassion; it reveals one aspect of the aretaic goodness of Tania. And we may claim that we value such acts because it is a good thing to have virtues and it is a good thing to act from these virtues. But we doubt whether this is the sort of appraisal of relevance with acts of love. There is the question, first, of whether love is a virtue. We may evade tackling this question head on by assuming that Tania is not what we would describe as a loving person; concerning love, she is unlike the way she is when it comes to compassion. Tania deeply loves Tully as she does her children. We would think well of her loving acts but *not* because they are revealing of her aretaic goodness. We would think well of them provided that she were commendable for them.

Arguably, there is more to the objection than what we have thus far acknowledged. Suppose Tania acts solely from compassion; her act is *wholly intrinsically motivated* in that she acts only for "the sake of compassion." Now the objection may proceed in this way. First, it may be submitted that we cannot be morally praise- or blameworthy for wholly intrinsically motivated actions. Since Tania acts wholly from compassion, she is not morally praiseworthy for so acting. Second, still, we would find persons morally meritorious in relation to their performance of such acts. We would value such acts as Tania's. Third, imagine that a person acts solely for the sake of love. Since she acts wholly from love, she is similarly, not praiseworthy from love's standpoint—she is not commendable—for that act; yet she is meritorious (because she acted wholly from love) *vis-à-vis* her performing that act. Hence, the thesis that what is valuable to us in behavior that love requires of us is essentially a function of our being commendable for this behavior is false. This objection requires careful development. We turn, first, to explaining its initial plank—that we cannot be morally appraisable for wholly intrinsically motivated actions—with which we agree.

8.3. WHOLLY INTRINSICALLY MOTIVATED ACTIONS AND MORAL APPRAISABILITY

Elsewhere, we have defended the view that moral appraisability—moral praise- and blameworthiness—requires conduct that causally derives partly from morally deontic beliefs. (Haji 1998, pp. 140–67) A morally deontic belief is a belief to the effect that something is morally right, wrong, or obligatory. More precisely, we have defended the following principle.

> *Appraisability:* One is morally praiseworthy (or blameworthy) for an action only if one performs the action, at least partly, in light of the belief that, in performing it, one is doing something morally permissible

> or obligatory, in the event of praiseworthiness (or morally wrong, in the event of blameworthiness).

In short, the principle at issue prescribes that appraisability requires action at least partly on the basis of relevant morally deontic beliefs (henceforth, the "at least partly" qualification will be suppressed but assumed). The primary thought underlying this principle is straightforward. Crucial to what we find commendable in an agent in instances in which the agent is morally praiseworthy for her behavior is her willingness (freely) to do right for right's sake; and, similarly, pivotal to what we find at fault in an agent in instances in which the agent is morally blameworthy is her willingness (freely) to do wrong. Action on the basis of the appropriate morally deontic belief manifests such willingness.

8.3.1. Arpaly's Challenge: Appraisability Without Morally Deontic Beliefs

Recently, Nomy Arpaly (2003) advanced a case that challenges *Appraisability*. She directs our attention to a well-known, key event in the adventures of Mark Twain's legendry character, Huckleberry Finn (1884/1985). The rudiments of the scenario building up to the event are familiar. Huck befriends Jim whom he helps escape from slavery. The deed deeply troubles his conscience. Like many others in his society, Huck thinks that aiding a slave to escape is tantamount to stealing, and that stealing is morally wrong. He also believes that one should be helpful and loyal to friends, but some things such as property rights outweigh loyalty to such acquaintances. Never doubting the mores of his society, he fails to find an excuse to help Jim escape. Having engaged in relevant deliberation, we may assume that Huck judges that, all things considered, he ought morally to turn in Jim. Yet, at the perfect opportunity to do what he believes is properly required of him, Huck freely and intentionally acts contrary to his best judgment; he discharges his moral obligation from weakness of will. Is Huck morally praiseworthy for this akratic deed? Arpaly sensibly proposes that the right answer depends on our reconstruction of Huckleberry's motives. On the interpretation she finds most plausible, Huck merits praise:

> On this interpretation, . . . during the time he spends with Jim, Huckleberry undergoes a perceptual shift. Even before meeting Jim, the way Huckleberry viscerally experienced black people was inconsistent with his "official" racist views. There are people who sport liberal views but cross the road when a person of a different race appears or feel profound disbelief when that person says something intelligent. Huckleberry, from the beginning, appears to be the mirror image of this sort of person: he is a deliberative racist and viscerally more of an egalitarian. But this discrepancy between Huckleberry's

> conscious views and his unconscious, unconsidered views and actions widens during the time he spends with Jim. Talking to Jim about his hopes and fears and interacting with him extensively, Huckleberry constantly perceives data (never deliberated upon) that amount to the message that Jim is a person, just like him. . . . [W]hen the opportunity comes to turn Jim in and Huckleberry experiences a strong reluctance to do so, his reluctance is to a large extent the result of the fact that he has come to see Jim as a person, even if his conscious mind has not yet come to reflective awareness of this perceptual shift. To the extent that Huckleberry is reluctant to turn Jim in because of Jim's personhood, he *is* acting for morally significant reasons. This is so even though he does not *know* or *believe* that these are the right reasons. The belief that what he does is moral need not even appear in Huckleberry's unconscious. . . . [M]y point is not simply that Huckleberry Finn does not have the belief that his action is moral on his mind while he acts, but that he does not have the belief that what he does is right *anywhere* in his head. . . . He is also unaware, or only dimly aware, of the fact that he is acting for these reasons in the first place. But he is acting for moral reasons all the same. . . . Huckleberry Finn, then, is not a bad boy who has accidentally done something good, but a good boy. (Arpaly 2003, pp. 76–77).

Arpaly's verdict, however, is controversial. According to her reconstruction, Huck acts on moral considerations, though he does not see them as such. Furthermore, Huck fails to act in light of the moral belief that his pertinent action is morally permissible or obligatory; Arpaly says that Huck has no such belief "*anywhere* in his head." However, without his germane action's causally arising from *any* such belief, skepticism regarding whether Huck is morally praiseworthy for this action is not uncalled for. Witness, for example, some of Frankfurt's remarks that bear on this concern:

> What counts in the assessment of a person's moral responsibility is not only what causes, reasons, or motives led to his action. It is also important to appreciate what sort of act he thought he was performing. A morally pertinent explanation of what a person has done must include an account of what he believed himself to be doing. . . . [Reconsider Green who was aware that he acted immorally—did moral wrong—in pursuit of selfish interests.] A full explanation of what Green did must provide more than just a statement that his motives were selfish. It must also report that he acted as he did because he cared more about attaining his selfish goals than he cared about avoiding immorality. This is relevant to his blameworthiness because it bears on what sort of action it was that he took himself to be performing. If he had performed the same act while believing it to be morally neutral, we would judge his conduct differently. (Frankfurt 2003, pp. 342–43)

As another representative of the opposing camp, when he assesses the possibility of an agent's being blameworthy for an action that is not morally wrong, Michael Zimmerman introduces cases where an admirable trait of an agent is not linked to the agent's belief about wrongdoing. In one case, Peter, moved by sympathy for his child, deliberately refrains from disciplining the child as he thinks he ought. In another of keen interest, Zimmerman says that Huck, who again moved by sympathy, refrains from thwarting Jim's bid to freedom as he thinks he (Huck) ought. Zimmerman remarks that some may venture that blaming these people is inappropriate. However, Zimmerman insists that we should not lose sight of the fact that, in these cases, the agent is indeed doing wrong from his perspective even if not in fact. He explains that Huck is to blame as long as it is accepted that he acted on the basis of the belief that he was doing wrong and that Huck satisfied other conditions of blameworthiness such as acting freely.[1]

As Arpaly presents it, Huck's case does not unequivocally support her verdict that Huck is praiseworthy and, hence, does not unequivocally speak against *Appraisability*. To advance the discussion, we provide an interpretation of Huck's case that appears to reinforce the verdict that Huck *is* deserving of moral praise despite lacking relevant morally deontic beliefs. The interpretation invokes wholly intrinsically motivated action. We proceed to argue that it is in fact doubtful whether one can be morally praise- or blameworthy—whether one can be morally appraisable—for wholly intrinsically motivated actions. This result *supports* the first plank of the objection against our thesis that what we find valuable in loving behavior is essentially tied to being commendable for that behavior. Despite this fact, we show that the objection fails.

8.3.2. An Alternative Reconstruction of Huck's Case

An *intrinsic desire* for something is a desire for that thing for its own sake or as an end. Mele explains that to desire something *wholly intrinsically* is "to desire it as an end and not also as a means to, as a constituent of, or as evidence of something else" (Mele 2003, p. 33). We can now give a more precise characterization of *wholly intrinsically motivated actions*. These are actions that causally (and non-deviantly) arise from wholly intrinsic desires; they are actions performed only for their own sakes (Mele 2003, p.71; 1992, p. 111). A relevant subset of such actions has traditionally been thought to be morally significant because members of this subset bear on the moral appraisal of agents. For example, Aristotle made it a necessary condition of being virtuous that an agent perform virtuous actions "for the sake of the acts themselves" (*Nicomachean Ethics,* 1144a8–20). If Russ displays kindness for its own sake in performing an action, his action causally stemming from a wholly intrinsic desire to display kindness, it would not be out of the ordinary to assume that Russ is morally praiseworthy for his display of kindness. It may be suggested that conceiving of Huck's

helping Jim as a wholly intrinsically motivated action will supply what is desired—Huck's case will, so augmented, provide credible support for a principle, discordant with *Appraisability,* that one can be morally praise- or blameworthy for intentional behavior not performed on the basis of pertinent morally deontic beliefs.

Elaborating, in acting as he does, imagine that Huck wishes to treat Jim as a person or, alternatively, Huck wishes to show loyalty to Jim. Suppose that treating Jim as a person (or showing loyalty) is, for Huck, an *end* and that he does not regard it as a means to any further end. His action would then be wholly intrinsically motivated. As Mele and Robert Audi have argued, standard Davidsonian explanations of such actions, in terms of a suitable want or conative element of the agent, paired with an apt belief, fail to hit the mark (Mele 1992, pp. 105–09; Audi 1986, pp. 543–44). Mele explains,

> In textbook instances of intentional action, an agent has a goal that he believes he can achieve by means of an action of a certain type. Al wishes to show Bob how much he appreciates his philosophical help over the years and he believes that an excellent way of doing this is to send Bob an autographed copy of his new book, writes the letter, and mails them to Bob. . . . Now suppose that Al has no ulterior motive for showing his appreciation to Bob. Suppose that showing his appreciation to Bob is, for Al, an *end* and that he does not also regard it as a means to some further end. . . . [His action is] *wholly intrinsically motivated*. . . . Ex hypothesi, it is not because Al believes his expressing his appreciation to Bob to be conducive to the achievement of a wanted item that he expresses his appreciation to Bob. To be sure, he does want to show his appreciation to Bob. But that want is linked, by a conduciveness belief, to his sending Bob the letter and autographed book, not to his showing his appreciation to Bob. In short, a belief of the sort called for in . . . [a standard Davidsonian account of an effective reason] is no part of anything that might count as the (or a) reason for which Al displayed his appreciation to Bob. . . . [T]he reason for which he performed this action was simply that he wanted to do this. If wants were brute forces wholly devoid of representational content, . . . [this proposal would be a *non sequitur*]. But wants are not like that at all. They do have representational content; for what is wanted is wanted under some conception or other. . . . Thus while the plan element of the reason for which Al sent the letter and book to Bob is provided by a belief that doing these things would be an excellent way of expressing his appreciation, the plan element of the reason for which Al showed his appreciation to Bob is provided by his *want* to do this—or more precisely, by the representational content of that want. (Mele 1992, pp. 106–10)

Reverting to Huck's case, it might initially be suggested that the reason for which Huck performs what has been stipulated to be a wholly intrinsically

motivated action—treating Jim as a person or showing loyalty—is constituted by a wholly intrinsic desire to perform an action with the attribute of *being a display of treating Jim as a person* and a belief that treating Jim in this way would have this attribute. However, as Mele clarifies (in the passage just cited), such a belief lacks an obvious explanatory function. More promising is to theorize that a wholly intrinsic desire to do something may itself—independently of a belief component—be understood as a reason for doing that thing (Mele 1992, pp. 104–12; 2003, pp. 71–72)

We shall then take it as a distinguishing feature of a wholly intrinsically motivated action that the wholly intrinsic desire from which it causally stems, independently of a belief component, constitutes its agent's reason for performing that action.

Entertain the view, consistent with Arpaly's interpretation of the case, that Huck's akratic act is wholly intrinsically motivated; the belief that he morally ought to treat Jim as a person does not play any role in its causal issuance. We may assume again with Arpaly that no "ought" belief of this or of a similar sort is "*anywhere* in" Huck's head. Still, one might think that there is something especially meritorious about Huck's act. The act stems from a desire solely for its own sake to treat Jim as a person and, thus, we have good reason to believe that, when all is said and done, Huck *is* morally to praise for this act. The principle that appraisability does not require action on the basis of morally deontic beliefs is, consequently, on firm ground.

Initial reflection on how wholly intrinsic desires may be acquired suggests that this line of reasoning, first, to the preliminary conclusion that Huck is morally praiseworthy, and then to the targeted conclusion that one can be appraisable without suitable morally deontic beliefs playing any role in the production of one's actions, may be on slippery footing. Wholly intrinsic desires for various things might be acquired on the basis of deliberation or other factors that do not, in any way, involve moral considerations. Ali believes that the preeminent policy for him to satisfy his desire to do best for himself in the long run includes cultivating and acting on wholly intrinsically motivating desires to display kindness to friends. On a certain occasion, he is kind to Jaya, his action stemming from a wholly intrinsic desire to display kindness on this occasion. Absent *any* morally deontic beliefs that assume any role in the action's genesis, it is *not* obvious whether Ali is morally praiseworthy for his wholly intrinsically motivated deed; it seems, in fact, that he is *not* praiseworthy. But if Ali is not praiseworthy for his deed, why think Huckleberry is praiseworthy for his, especially when Huck, unlike Ali, takes himself to be doing intentional moral wrong?

Prior to arguing for the view that one cannot in fact be appraisable for wholly intrinsically motivated actions, preliminary remarks are in order. First, it would be presumptuous to suppose that Arpaly is committed to the position that Huck's germane act is wholly intrinsically motivated. However, the modification proposed lends a plausible gloss to Huck's case

and, as suggested, it may at least initially help to bolster support for the contentious verdict that Huck is morally to praise for his akratic deed. In any event, attention is largely confined to this reading of Huck's case. We argue that, given a reasonable action explanation of Huck's deed, Huck is not morally praiseworthy for this deed. This will serve to deflect *one* line of argument to the contrary verdict, and, hence, circumvent a seemingly compelling pathway of reasoning for the principle that appraisability does not require action on the basis of morally deontic beliefs. Second, Arpaly does *not* defend the view that one can be appraisable for wholly intrinsically motivated actions and we do not wish to leave the impression that she does. We focus on such actions, partly, to address two other objectives: defending the moral that it is not evident whether any agent *can* be morally appraisable for a wholly intrinsically motivated action, and inquiring into how to assess agents, when the assessment is of a moral variety, in relation to their performance of such actions.

8.3.3. Appraisability for Wholly Intrinsically Motivated Actions?

A feature of wholly intrinsically motivated action warrants reemphasis. Suppose Jim shows appreciation to Huck and that this act of his is wholly intrinsically motivated. Then it cannot *also* be the case that, for instance, his act arises, in part from the belief, if he has it, that showing appreciation to Huck on this occasion is morally right. In contrast, an action arising from an intrinsic desire, but not a *wholly* intrinsic one, to show appreciation may be generated partly on the basis of this sort of belief.

Consider a progression of cases that casts doubt on whether one can be appraisable for a wholly intrinsically motivated action. In the first, imagine that Alia-1 is a cognitively sophisticated alien—more or less as complex in this respect as any normal, mentally healthy, adult human being—who has a rich psychological life. Alia-1, however, is devoid of moral concepts. Though she assesses behavior normatively, none of these evaluations is a moral one. In particular, she has no grasp of the categories of moral right, wrong, and obligation. Suppose that on some occasion, she freely and intentionally performs an action, such as helping a friend, because she correctly believes that the action is prudentially best and she desires to do best for herself. Even if this action coincides with the morally right thing for her to do, she is not morally praise- or blameworthy for performing it, or for that matter, for any other action. Simply put, she is an amoral agent.

This verdict accords with two prominent views concerning what judgments of moral responsibility are about. The first is that to be morally responsible just is to be the appropriate object of what Strawson has called the "reactive attitudes," such as gratitude, resentment, indignation, and the like. Strawson explains that it matters to us whether the actions of other people "reflect attitudes towards us of good will, affection, or esteem on the one hand or contempt, indifference, or malevolence on the

other" (P. Strawson 1962/1982, p. 63). The reactive attitudes are "natural human reactions to the good or ill will or indifference of others towards us as displayed in *their* attitudes and actions" (P. Strawson 1962/1982, p. 67); and they express "the demand for the manifestation of a reasonable degree of good will or regard, on the part of others, not simply towards oneself, but towards all those on whose behalf moral indignation may be felt" (P. Strawson 1962/1982, p. 71). On this view, a person devoid of moral concepts is exempt from responsibility because this person's conduct expresses neither an attitude of moral ill will nor one of moral good will toward anyone. If being responsible is to be understood in terms of the stance of holding responsible, and if what it is to hold a person morally responsible for wrong conduct is nothing more than the propensity toward, or the sustaining of, a morally reactive attitude of disapprobation, and, further, the disapprobation is in response to the perceived attitude of *moral ill will* in the conduct of this person, then Alia-1, on the Strawsonian view, fails to qualify as an appropriate candidate for the moral reactive attitudes.

On the second view of what judgments of responsibility are about, to be morally responsible is to be such that one's moral standing or record as a person is affected by some episode in, or aspect of, one's life. As M. J. Zimmerman explains, the first and the second views are allied, the difference between the two being that, "whereas the former identifies responsibility with susceptibility to the reactive attitudes, the latter identifies responsibility with *that in virtue of which* one is susceptible to the reactive attitudes" (M. J. Zimmerman 2002, esp. sec. 1). On this second view, when a person is praiseworthy, her moral standing has been enhanced in virtue of some episode in her life; when blameworthy, her moral standing has been diminished. As Alia-1 fails to have any *moral* record or ledger, no episode in her life can enhance or diminish her *moral* standing.

In a second case, Alia-2 does have a grasp of moral concepts. She understands, for instance, what it is for acts to be morally obligatory or supererogatory. She does not care about morality in the sense that morally deontic beliefs—beliefs about moral right, wrong, or obligation—are *typically* not any part of the psychological antecedents that move her to action. She sometimes does what is morally required of her akratically—as we shall abridge, she sometimes performs "morally akratic acts." However when, like Alia-1, she helps her friend, she does not succumb to morality out of weakness of will. Further, when she helps, morally deontic beliefs are not part of the etiology of this action, and she does not regard the proximal desire that moves her to action to be a moral consideration. It seems that, in this case, Alia-2 is not morally praiseworthy for her deed. This assessment may be supported, first, by noting that responsibility depends on the *actual sequence of events* that leads to action. Despite their differences concerning their grasp of moral concepts, and the fact that Alia-2 occasionally performs morally akratic acts, there is

nothing to preclude its being the case that the actual sequence of events that culminates in Alia-1's deed is type-identical (or near type-identical) to the actual sequence of events that eventuates in Alia-2's.

Second, a principle concerning the irrelevance of facts to the explanation of behavior also supports the judgment that Alia-2 is not praiseworthy for the pertinent act. Frequently, in response to why someone acted as she did, it suffices for the purposes at hand to give what we can dub "narrow" action explanations of intentional action in terms of prior proximal causes of the action where these causes may be suitable desire/belief pairs or, perhaps, simply desires as with wholly intrinsically motivated action. Narrow explanations, though, will not provide a fuller account of the behavior when what is sought is a particular kind of illumination or more probing detail. For example, we may give a narrow reasons explanation of why Huck performed the action that he did which will invoke appropriate prior psychological states of Huck or events involving Huck. However, this sort of explanation will not account for why Huck's deed is an akratic one. Much more will be required to come to an appreciably full understanding of his akratic behavior.

When we ask for fuller explanations, we are customarily seeking a more comprehensive explanation, over and above a merely narrow one, for a particular feature of the behavior of interest. We may, for example, want to know why the agent acts akratically, or from self-deception, or why, in doing as he did, the agent is susceptible to moral appraisals of various sorts—why, for instance, is the agent morally praiseworthy? We can say that a fact is relevant to a fuller explanation of why an agent did what he did if the fuller explanation of this behavior requires appeal to this fact. As an illustration, in a highly promising account of akratic action, the motivational strength of the agent's desire to perform the akratic action is misaligned with the agent's evaluative assessment of what the agent desires; this fact is relevant, on this account, to understanding akratic deeds. (Mele 1987) We can now, in coarsely chiseled strokes, cast the principle of explanatory irrelevance in this way:

> *The Principle of Explanatory Irrelevance:* If a fact is not relevant to a fuller explanation of the why an agent did what he did, then this fact has no bearing on the feature of interest of the agent's action or of the agent in relation to the action, such as being appraisable for the action.

Reverting to Alia-2's case, the fact that Alia-2 sometimes performs morally akratic acts, or the fact that she sometimes acts (partly) on the basis of morally deontic beliefs, or the fact that she has a grasp of moral concepts, is not relevant to a fuller explanation of why Alia-2, *on the occasion of interest,* helps her friend. Hence, the principle yields the result that these facts have no bearing on whether Alia-2 is morally praiseworthy for helping her friend. Facts such as these are the only pertinent ones that

distinguish Alia-2's case from Alia-1's. Since Alia-1 is not praiseworthy for the relevant deed, we may conclude that Alia-2 is not praiseworthy for her deed either.

Now consider Alia-3 who frequently, though not always, is moved by moral considerations: she often acts in light of morally deontic beliefs and on desires whose content may be deemed moral in the sense in which Arpaly seems to think Huck's germane desires—his desires to treat Jim as a person and to show loyalty to him—implicated in his helping Jim escape are moral. Alia-3 helps Ahmed and this act of hers is wholly intrinsically motivated. Assume that this act is morally obligatory for Alia-3, but since it is *wholly* intrinsically motivated, it does not causally arise, even, partly, from any morally deontic belief of hers. Imagine that Alia-3 has a generic desire to do whatever is morally required of her whenever she acts. But suppose that this desire, too, on this occasion, is on the sidelines: it is causally inefficacious, failing to play any role whatsoever in the production of her action. It does not, for example, generate, on the basis of practical deliberation, the belief that she morally ought to help Ahmed or does not, in any other way, enter into the etiology of her helping Ahmed. Focusing on the actual sequence of events that culminates in her relevant action, Alia-3's case is no different along this dimension than the first two. One would be hard pressed to suppose that Alia-3 expresses moral good will in her conduct. (Similarly, of course, it would be unreasonable to suppose that her conduct expresses moral ill will.) Nor would it be credible to suppose that her helping Ahmed enhances her moral standing as a person. This is primarily because she fails to perform the action in light of *any* belief that what she intends is morally right or obligatory.

In addition, the fact that helping Ahmed is obligatory for her is explanatorily irrelevant to her helping Ahmed. We may even suppose that Alia-3 believes that she morally ought to help Ahmed, but this belief too, given that her act is wholly intrinsically motivated, is on the sidelines. Then this fact—that she takes helping Ahmed to be obligatory for her—will not figure in an explanation of why she helps Ahmed, anymore than will the fact that she has a grasp of moral concepts, or the fact that she frequently acts on what she takes to be moral considerations, or the fact that she has a generic desire to do what is morally required of her. The principle of explanatory irrelevance yields the result that these facts do not bear on Alia-3's being appraisable for helping Ahmed. Again, it is facts of this sort that distinguish the third case from the first and the second. Consequently, if neither Alia-1 in the first case nor Alia-2 in the second case is morally praiseworthy for her pertinent behavior, it is difficult to see why Alia-3 should be praiseworthy for her germane action in the third.

In the story as Arpaly introduces it, Huck acts on a moral reason though he does not see this reason as a moral reason or consideration. In Alia-1's case, there is a sense in which Alia-1 acts on a moral reason. Assume that helping is morally obligatory for her, and because it is morally obligatory,

we can say that there is a moral reason for her to help. There is such a reason even though she has no understanding of morally deontic considerations. There being a moral reason of this sort for her to help should not alter the assessment that she is not morally praiseworthy for helping. We may assume that there is a moral reason of this kind in the second and third cases as well. Alia-3's act, unlike Alia-2's, is wholly intrinsically motivated. Although it is not done for any further reason, her intentional act is, as Mele (2003, p. 72) very plausibly suggests, done *for* a reason: the reason is constituted by her intrinsic desire to help Ahmed. Alia-3, as we have imagined, believes that she morally ought to help Ahmed on the particular occasion, but she can entertain this belief consistently with its not being the case that she regards the reason for which she helps Ahmed—the wholly intrinsic desire—as a moral consideration in favor of helping Ahmed. However, then Alia-3's case will, once again, not differ from Alia-1's in the respect that neither of the agents sees the reason for which she acts as a moral reason. If all other pertinent considerations remain the same in the two cases, we should continue to regard Alia-3 as not being praiseworthy for helping Ahmed if we regard Alia-1 as not meriting praise for her deed.

There may well be another sense in which Huck acts on a moral reason in addition to the thin sense in which what Huck does is morally obligatory. Let us assume that this alternative sense of acting on a moral reason is closely pegged to the content of the pertinent desire of Huck's. The notion of a desire's content, though, is ambiguous. On the first reading, what is wanted is wanted under some description or another; so wants have representational content. On the second reading, it is not so much how the agent represents the wanted item that is pertinent—representational content is not at the fore—but rather what is germane is, roughly, what the desire is *really* about. The focus here is on defining or characteristic features of the object of the desire even when these features are not salient in the agent's conception of the object because the agent may be fundamentally mistaken about the nature of this object. The reconstruction of Huck's scenario with which we began suggests that, in not turning Jim in, Huck acts on the desire to treat Jim as a person—a human being (Arpaly 2003, p. 77); presumably, Huck's representational content of this desire—if it even makes sense to speak of its representational content because the desire is unconscious—differs from content so conceptualized which we may refer to as "plain content." In either sense of content, we must be careful to construe Huck's case as a case in which the desire is free of any content, representational or plain, that would implicate Huck's having a belief that he ought morally, or that it is morally permissible for him to help Jim escape, or his taking himself to have a moral reason to help Jim escape. This is in keeping with the specifications of the case that no pertinent morally deontic beliefs are *anywhere* in Huck's head—he does not, for example, act on the basis of a belief to do the right thing—and Huck does not regard the reasons for which he helps Jim as moral reasons.

On the one hand, suppose it is plain content that is central to explicating this alternative conception of there being a moral reason to do or refrain from doing something. If possession of such content does not even require that one be aware that the desire in question is morally relevant or qualifies as a moral consideration or reason, then nothing prevents Alia-1's desire to help from having "plain moral content." The fact, if it is one, that her desire has such content should not alter our verdict that she is not morally praiseworthy for helping.[2] Then too, all else remaining the same, we should not be swayed by the contrary verdict when it comes to Alia-2's, or for that matter, Alia-3's helping. Regarding the latter, her wholly intrinsic desire to help Ahmed may have plain moral content; but why should this make any difference to our assessment of whether she is praiseworthy when there is no shift in our assessment of whether Alia-1 is praiseworthy on the supposition that Alia-1's desire to help has such content?

On the other hand, assume that it is representational content that is of concern. We must be careful to remember that our agents—Huck and the Alia sisterhood—cannot represent what is desired under a conception that their pertinent action is morally obligatory or counts as a moral consideration. Alia-1's desire to help obviously meets this constraint because she is an amoral agent. We may suppose that she represents the item that she desires under some such description as "wanting to help a friend" or "wanting to help a friend as a way of expressing thanks." Alia-3's desire to help Ahmed is wholly intrinsic. The representational element of this desire is just her wanting to help Ahmed for its own sake. Now again, compare the actual sequences of events that culminate in action in these two cases. There is no bar to supposing that they *can* be type-identical. Then the lesson is clear: if Alia-1 is not praiseworthy for helping, Alia-3 should not be either.

Supplement Alia-3's scenario in a fashion adequate to Alia-3's acting out of friendship but *not* out of any concern—specific or generic—to do the morally right or morally obligatory thing. In helping Ahmed, assume that Alia-3 is doing what is morally required of her. Further, in so acting, assume that she believes that she is morally required to act as she does on this occasion and, more generally, she believes that she morally ought to perform actions of this sort under similar circumstances. However, assume, again, that she is not acting even partly on the basis of these beliefs but she is acting (solely) out of friendship.[3] Here too, it seems that though Alia-3 may be non-morally but normatively praiseworthy for her action—she does, after all, act out of friendship—she is not deserving of *moral* praise. (In the terminology that we introduced in the last chapter (Section 7.2.2.) it is false that Alia-3 acts narrowly from moral duty.) The causal history of her action is relevantly similar to the causal histories of the pertinent actions of Alia-1 in the first case and of Alia-2 in the second. We may admit that there is a sense in which Alia-3 does the right thing for moral reasons if acting out of friendship constitutes a moral reason, but

this sort of moral reason is insufficient for moral praiseworthiness. Were it sufficient, Alia-1 would be morally praiseworthy in a scenario in which she acted out of friendship.

This sequence of cases forcefully suggests that a vital element that counts in the assessment of whether an agent is morally praiseworthy—or more generally, morally appraisable—for an action is whether the belief that what she was doing is right, obligatory, or wrong plays a suitable role in the causal genesis of the action. In particular, what is salient is whether the agent acted (at least partly) *on the basis* of such a belief, occurrent or dispositional. For in the absence of acting in light of such a belief, it seems that the agent can express neither moral good will nor moral ill will in her conduct or that her moral record as a person—the record pertinent to praise- or blameworthiness—can be neither enhanced nor diminished.

Reverting to Huckleberry's scenario, if Huck's treating Jim as a person is a wholly intrinsically motivated action, Huck is not morally praiseworthy for this deed. It may be that Huck desires to treat Jim as a person (or show loyalty to him) and believes that he will achieve this end by helping Jim escape. In this latter variation, provided that his action is not based, even partly, on the belief—occurrent or dispositional—that he is doing something right or obligatory in acting as he intends, again, there is little reason to suppose that he is deserving of moral praise. We can, consequently, conclude that one line of argument for the view that one can be morally appraisable for an action even if the action does not stem from morally deontic beliefs is not sound. In the absence of other telling considerations we may, thus, continue to adhere to principle Appraisability that moral appraisability requires action performed on the basis of morally deontic beliefs.

8.3.4. On the Moral Assessment of Wholly Intrinsically Motivated Actions

We can now draw a general lesson concerning appraisability for wholly intrinsically motivated action. As explained, such actions causally (and non-deviantly) issue from wholly intrinsic desires. Even if the agent has pertinent moral beliefs concerning whether the intrinsically motivated action performed is morally right, obligatory, or wrong, or a generic moral belief that she ought always to act as morality requires, these beliefs are on the sidelines: she does not act, even partly, on the basis of them. It seems, then, that one cannot be appraisable for *wholly* intrinsically motivated actions.

What then, though, of Immanuel Kant's suggestion that one is praiseworthy for an action (in Kant's terminology, an action has moral worth) *only if* its agent performs it for the sake of duty? (I. Kant 1795/1964, pp. 65–66) Can it not be that an action is, so to speak, wholly intrinsically motivated by moral duty—by moral obligation? The Kantian thesis is susceptible to two interpretations. On the first, the thesis is that an action has moral worth (that is, its agent is praiseworthy for the action) only if its agent

performs the action for the sake of duty. On the second, the claim is that an action has moral worth only if its agent performs it *solely* for the sake of duty. We agree that the second interpretation, as some believe, is mistaken, but this agreement is not of fundamental import to our concerns.[4] We propose that on a reasonable analysis one acts from duty—either from duty itself or when motivation to act from duty is concurrent with some other non-primary motivation—only if one acts on the basis of the belief that one's act is morally obligatory (or morally right). Briefly put, one acts from duty only if one acts in light of appropriate deontic beliefs; one does right for right's sake. And if this is so, acting solely from duty does not jar with the position that one cannot be appraisable for *wholly* intrinsically motivated action.

Can there, though, be cases in which an agent's intentional act is done for a reason *exhaustively* constituted by her desire to do moral right, and at that, such a desire under that description? In such cases, the agent would want to do moral right but would fail to do what she does on the basis of the belief that what she is doing is morally right. Such cases appear to be either incoherent or if not incoherent, bizarre. Suffice it to say that if there are such wholly intrinsically motivated actions, and if some deem it appropriate to refer to such actions as actions stemming wholly from moral duty, then one cannot be appraisable for such actions.

Yet against the view that wholly intrinsically motivated action cannot be action for which one is appraisable, it may be rejoined that, regarding assessments of praise- or blameworthiness, the causal genesis of desires, including the genesis of wholly intrinsic desires, matters at least in this respect: if these desires are acquired on the basis of apt moral deliberation, then the agent could be appraisable for actions that issue from these desires even if the actions, when performed, are not performed in light of appropriate moral beliefs. Suppose that, given his upbringing, Sam acquires the generic desire to do whatever is morally required of him on each occasion, and that this desire, in conjunction with appropriate deliberation, gives rise to the generic desire to display kindness to friends. The generic desire, supplemented with relevant beliefs, in turn, generates in him the wholly intrinsic desire to display kindness to Al now. Should Sam not be morally praiseworthy for his wholly intrinsically motivated action of displaying kindness to Al now even if he fails now to act in light of the moral belief that what he is doing now is morally right or obligatory? The answer, we suggest, is that he should not be praiseworthy. We may acquire desires as a result of deliberation or on the basis of other considerations that fail to involve beliefs of moral right, wrong, or obligation and, yet, when we satisfy such desires we may act partly on the basis of such beliefs. In such instances, we could well be morally appraisable for our actions. It is also true that we may acquire desires on the basis of factors that do involve beliefs of moral right, wrong, or obligation, as Sam does, and yet, when we act on them, not act in light of any such beliefs. Why, then, should we

be appraisable for the germane actions that issue from these desires? Moral praise or blame is *not* so easily merited. We need not, of course, deny that it is a good thing that Sam acted as he did or that he may have acted from the virtue of kindness. However, these varieties of moral assessment differ from appraisability. Conflating them with appraisability simply masks the complexity of the moral life. Nor need we deny that Sam may well be to praise for *acquiring* the desires. But it does not follow that Sam is to praise for actions issuing from such desires.

Suppose Alia helps Mia but not only out of a concern for Mia's well-being or happiness but also partly on the basis of her moral belief that she ought to help Mia. That she believes she morally ought to help Mia is a non-trivial contribution to Alia's decision to help Mia. Suppose, in contrast, that Alia* helps Mia solely out of concern for the well-being of Mia, her action issuing from a wholly intrinsic desire. No moral belief of right or obligation is part of the motivational mix that gives rise to her act. Is Alia* a better moral agent—more morally perfect as some would have us believe—than Alia?[5] This way of posing the question, it seems, is not enlightening. It is not true, for instance, that both agents are morally appraisable. Whereas Alia *is* morally praiseworthy for helping Mia, assuming other conditions of responsibility are satisfied, Alia* is not. For Alia* fails to act on the basis of the belief that she is doing something that is morally right (or obligatory), despite, perhaps, believing that what she is doing is morally right (or obligatory). Alia* though, does we assume, do what is morally right and she may well act from virtue. It is more illuminating, at least as far as moral assessment is concerned, to discern that an agent can, in principle, be open to different sorts of moral appraisal and score high on some scales and low on others. Comparatively, Alia does well regarding moral appraisability but may depending on how the details unravel, score low on the scale of aretaic assessment. Whether one agent is more morally perfect than another is, given the variety of moral appraisals, a complicated matter and may, in any event, be entirely beside the point if one has a sound picture of how the agent fares on the different scales of assessment.

We need to introduce a final set of considerations before reverting to the objection against our thesis that the value of loving behavior for us resides in our being commendable for that behavior. We previously suggested that some actions that are wholly intrinsically motivated are especially morally meritorious. They are so, it may be thought, in that they reveal something particularly morally worthy about their agents. Is this in fact true? It is worth pursuing the idea that at least some actions that are wholly intrinsically motivated—such as Alia's displaying justice solely for its own sake—are actions performed from virtue. Thus, such actions will exemplify moral worth insofar as such worth is taken to express the (aretaic) goodness of their agents. What precisely is involved in acting from virtue is enormously complex. In comparing the arts and the virtues, Aristotle outlines the following conditions:

> [T]he products of the arts have their goodness in themselves, so that it is enough that they should have a certain character, but if the acts that are in accordance with the virtues have themselves a certain character it does not follow that they are done justly or temperately. The agent must also be in a certain condition when he does them; in the first place, he must have knowledge, secondly he must choose the acts, and choose them for their own sakes, and thirdly his action must proceed from a firm and unchangeable character. (Aristotle, *Nicomachean Ethics,* 1105a29ff)

Aristotle's first point seems to be that whereas an act may be in accord with justice, it would not express the virtue of justice if, when performed, it was not performed in the right way because the agent was not in the right state. Of particular interest to our concerns are the recommendations that if the agent is to be in the right state to act from, for example, justice, he must know that he is so acting, and the virtue (or virtues) at issue must appropriately motivate it. The motivation condition requires that acting, for example, from justice, calls for the agent's acting justly for its own sake; we may take Aristotle to be suggesting that such action requires being motivated by at least an intrinsic desire to display, judge, or act justly. (We set aside the two other proposals in the passage: acting from virtue requires that the pertinent action be decided upon and that it stem from the appropriate character trait.)[6]

The knowledge condition seems to entail that, if Alia acts from justice, she is aware or understands that her act is appropriately connected with justice. Stronger and weaker renditions of this awareness condition are possible. A strong reading would require that, when for instance, an agent acts from justice, she is aware (or at least believes with justification) that her act issues from motivation that incorporates a suitable concept of justice. Roughly, the agent conceives of the pertinent act as an act of justice; she is aware or justifiably believes that the act is performed from a desire to achieve that virtue. A weak reading might require only that the agent believe that her action issues from a desire to achieve something that has at least some of the primary features of justice; acting justly would, in this probably more typical case, involve sensitivity to features that make acts just.[7] Alia, for example, may act from a desire to give equal compensation to male and female employees for comparable work; she aims to achieve justice even though the concept of justice itself does not figure in her motivation to act as she does. The point of significance for our purpose resides in Aristotle's suggestion that without an adequate understanding, strong or weak, of one's action's being suitably connected with the virtue at issue, one's act cannot express that virtue. This has the consequence, as Aristotle sees it, that one would not then be suitably praiseworthy for the act.

The knowledge component of Aristotle's analysis of acting from virtue seems to imply that, if Huck is (strongly or weakly) unaware that, in not

turning Jim in, his intentional omission issues from a moral duty or virtue, then Huck cannot be morally praiseworthy for this omission. In addition, Aristotle's knowledge condition forcefully suggests or even lends support to the principle that appraisability requires action on the basis of morally deontic beliefs. If Augustine does not believe that his act of giving alms to the poor is even partly performed from motivation, a doxastic element of which is that his giving alms on this occasion is morally required of him, it would seem that he is not *morally* praiseworthy for giving alms.

Turning next to the intrinsic motivation condition of the analysis, Aristotle seems to allow both for an action from virtue's being wholly intrinsically motivated, as when Alia performs an act of justice solely for the sake of justice, and for "mixed" motivation cases in which an action can express a virtue even if other items, besides elements of the virtue, play a causal role in giving rise to the action. The former "pure case" imparts credibility to the thought that an appropriate subset of wholly intrinsically motivated actions are especially morally revelatory of persons—the agent has acted from virtue, say the virtue of justice or kindness—and in so doing, has revealed her aretaic goodness. She has revealed such goodness (perhaps) even if she is not free with respect to being just or kind. The mixed case has a bearing on moral appraisability. An intrinsically motivated action may have motivational components other than its intrinsic desire; the action may stem partly from a relevant belief of moral right, wrong, or obligation. Ahmed may act from courageousness—his act issuing from an intrinsic desire to realize the virtue—but the act may also be based partly on the belief that he is morally obligated to act as he does. In this instance, Ahmed's act would have moral worth and he may also be morally praiseworthy for it. If causally produced in the *absence* of the appropriate influence of any morally deontic beliefs, Ahmed's act of courageousness would still exhibit moral worth. He would not be morally praiseworthy for this act but there is no reason to suppose that he would not be positively evaluable on the aretaic scale. Again to belabor the point, the aretaic goodness of persons is one sort of moral assessment; appraisability is a different sort. What may be especially morally meritorious about certain wholly intrinsically motivated actions is that they stem from virtue and thus exhibit moral worth—they are revealing of aretaic goodness.

8.4. THE OBJECTION RECONSIDERED

The objection against the thesis that the value we find in loving behavior is essentially a function of agents being commendable for loving behavior invokes wholly intrinsically motivated action. The first step of the objection denies that we can be morally appraisable for wholly intrinsically motivated action. We agree. Tania is not morally praiseworthy for helping the other if she acts solely from compassion. The second step says that despite

its not being the case that we are appraisable for pertinent wholly intrinsically motivated action, we find persons morally meritorious in relation to their performance of such acts. We value such acts. Once again, we agree. We stress that what we find morally meritorious in persons vis-à-vis performance of such acts, when they are meritorious in relation to such acts, is that these acts are expressive of their aretaic goodness. The third step in the objection introduces a case in which a person acts solely for the sake of love. It is recommended that since this person acts wholly from love, she is not commendable for that act; yet she is meritorious (from love's standpoint) for that act. We value such acts. Hence, the thesis that what is valuable to us in behavior that love requires of us is essentially a function of our being commendable for this behavior is false. This third step demands close scrutiny.

To begin, we stress that if Tania acts wholly from compassion, what we find meritorious about Tania in relation to her compassionate act is that her act expresses a virtue; it reveals Tania's aretaic goodness. If Marian though, acts solely from love, her loving act does not reveal her aretaic goodness. It is after all, doubtful whether love is a virtue. So, as it stands, the third step is not compelling.

However, there is still something in the third step that merits further discussion. We have proposed that commendability is a species of praiseworthiness—it is praiseworthiness from the point of view of love. It is one sort of normative praiseworthiness. We have claimed that a person is morally praiseworthy for performing an action only if she does it partly on the basis of the belief that the action is morally obligatory or right; moral praiseworthiness requires conduct on the basis of appropriate morally deontic beliefs. Then it would seem that there should be an analogous doxastic requirement for commendability as well. One is commendable for some action only if the loving action issues appropriately from the belief that that is what love, on the occasion, requires (or permits) of the agent. If what is of value in loving behavior is essentially tied to commendability for such behavior then, given the doxastic requirement of commendability, it follows that what is of value in loving behavior presupposes that the agent performs the loving action at least partly on the basis of the belief—occurrent or dispositional—that love requires (or it is permissible for her from love's standpoint) that she perform this action. However, one may well balk at this result. Reflecting on some of Philip Pettit's views on love will bring out the concern (Pettit 1997).

Pettit argues that love is not a virtue because love does not display the same explanatory-justificatory "structure" associated with the virtues such as kindness or fairness. Behaving in a kind way can be invoked both to explain and to justify a person's behaving in that way. The fact that the option is kind will serve to justify the choice of it. And the fact that someone believes that the option is kind or more typically, the fact that someone believes that the option has features that in the context, qualify it as being

kind, will serve to explain the choice; it will serve to explain what moves the agent to make that choice (Pettit 1997, pp. 154–55). Love though, lacks the dual explanatory-justificatory role of kindness or the other virtues:

> The fact that I love someone may serve to justify my treating her in a certain, say, partial or self-sacrificing way. But it is doubtful whether I could claim to be properly a lover, if it was my recognition of the fact of loving her—or my recognition of a realiser of that fact—which explained my action: if all that needed to be said in explaining how I behaved was that I saw I loved her or saw I bore a relation to her which, as it happens, means that I loved her. . . . This may seem too quick. Perhaps I am moved in love, as I am moved in kindness, by a recognition that the acts I choose have features, whether or not I see them in this way, that make them loving acts. Perhaps love and kindness show their similarity at the level of acts: kindness involves a sensitivity to features that make acts kind, love a sensitivity to features that make acts loving. . . . But a little reflection reveals a fatal weakness in this suggestion. Someone may be sensitive to features that make acts loving in relation to someone, not because of being truly in love, but rather because of being committed to behaving in a loving way: not because of a lover's commitment, as we might put it, but rather because of a commitment to love. . . . The characteristic explanation of a lover's behaviour towards a beloved is not the recognition of the fact of loving her, nor the recognition of the presence of any related features, but rather the fact of loving itself. . . . Suppose that my behaviour was not to be explained in this characteristic way but rather in the manner of a virtue like fairness or kindness. Suppose, for example, that I tried to keep note of the person's birthday, that I gave freely of my time to help her, and so on, because of registering in each case that this was someone I loved: because of registering this and not, as we would say, because I loved her. In that case, I might be praised for my moral determination to honour the relationship but I could not be said, without qualification, to be acting out of love. To act out of love, as we might put it, is to be moved by love and not by the recognition of being in love. (Pettit 1997, pp. 155–56)

These remarks of Petitt's suggest a revision of the third step of the objection under consideration. The value we find in lovable acts is essentially a function of these acts *manifesting* love. Further remarks of Petitt's suggest that an act can manifest love—one can act out of love as Pettit puts it—without acting in light of the belief that love requires that the lovable act be performed. Pettit proposes that what is necessary and sufficient to act out of love (and, hence, what is necessary and sufficient for an act to manifest love) is motivation to perform the act by a belief or consideration that is "rigidly individualized": a consideration of the sort that there is "no way of knowing

exactly what the content of the consideration is—no way of understanding it fully—without grasping who the particular [beloved] is" (Pettit 1997, p. 158). The consideration must identify the beloved essentially by way of a name or demonstrative such as "This, my friend, is in need" or "Tania is in need." The reason is that in acting out of love, what moves one to so act, must essentially involve the beloved; the beloved is the primary focus of the motivation and not, for example, some feature of the agent, such as being loved by, or being in a loving relationship with, the agent:

> [W]hen love is manifested in the canonical way, when an agent displays a commitment to a beloved by acting out of love, then the reason that moves the agent has to be rigidly individualized in favour of the beloved. It has to be a reason in which the beloved figures as an essential component, whether by courtesy of a name or demonstrative or whatever. And it has to be a reason that moves the lover, at least in part, by virtue of involving the beloved in that way. (Pettit 1997, pp. 158–59, note omitted)

In sum, building on these views of Pettit, one may attempt to impugn the thesis that the value we find in lovable acts is essentially a function of commendability for these acts by suggesting that the value is essentially a function of the pertinent acts manifesting love, or alternatively, of the pertinent agents acting out of love. Further, acting out of love is not mediated by a belief on the agent's part that love requires the agent to perform the lovable act. If one's motivation to act included a belief of this sort, the motivation would not be "focused" on the beloved; it would feature some merely accidental property, *being such that love requires that the agent perform the pertinent act*.

We believe however, that this attempt to undermine the relevant thesis fails. First, typically, when a person acts out of love, we presume that the person *is* commendable for the lovable act. If the person were *not* so praiseworthy, we would not think of the act as manifesting love. The act would be a "lovable act" only in that it satisfies a requirement of love. Perhaps one might be congenial to the view that, typically, if a person acts out of love, the person is commendable for the act but disagree that commendability requires the relevant belief. We do not need to settle this dispute because the concession leaves intact the thesis under scrutiny.

Second, ponder the claim that what is necessary and sufficient to act out of love is motivation by a belief that is rigidly individualized; it is this sort of belief that moves the agent to perform the lovable action when she acts out of love. Suppose Natasha gives up one of her kidneys to save her daughter's life, and she acts out of love when she does so. The belief that this is Tania, and Tania needs the kidney, cannot, it seems, be *sufficient* motivation *to act out of love*. (Let's simply sidestep the issue of whether beliefs, on their own can be motivating. Should one be skeptical about this

view, assume that Natasha has a pertinent desire, say, one to help needy Tania). Nor it seems, could the belief whose propositional content is "This is Tania, my daughter, who needs a kidney to survive" be sufficient. For if a belief of this sort were sufficient to act out of love, what would we say of a case in which Natasha has this belief, this belief is partly what moves her to action, but it is a case in which Natasha acts from moral or prudential considerations? Again, imagine that she acts partly on the basis of the belief that she is morally obligated to give up the kidney. One may, of course, insist that we separate such "mixed" motivation cases from ones in which the relevant belief is the sole motivator. But we do not see how this helps at all. If one insists that what solely moves Natasha to perform the allegedly lovable act is the belief: "This is Tania, my daughter, who needs a kidney," what is the basis for claiming that Natasha acts out of love in preference to acting out of some other consideration? If she *regards* the relation as a loving one, and this awareness figures partly in moving her to act, there are then grounds to suppose that she acts *out of love* rather than out of some other normative consideration. Thus, we do not see how a rigidly individualized belief can be *sufficient* motivation for acting out of love, anymore than that it can be *sufficient* motivation for acting out of say, moral duty, or religious conviction.

We conclude that the foregoing considerations do not dislodge the thesis that the value of lovable behavior is essentially a function of agents being commendable for the behavior. Behavior merely in accordance with the requirements of love but for which its agent is not commendable is not genuinely loving behavior.

8.5. ACTING FROM LOVE VERSUS ACTING FROM DUTY

Setting aside wholly intrinsically motivated actions, we have proposed that a person can at times, perform an act out of love and hence, be commendable for that act but not perform it at that time out of (moral) duty and hence, not be morally praiseworthy for it. However, this view has not gone unchallenged. In the remainder of this chapter, we summarize some of David Velleman's central theses on acting from love and acting from duty that implies that our proposal is mistaken. We dispute elements of his views to defend our position. The discussion will shed further light on the distinction between acting from love and acting from duty.

The argument implicit in Velleman's views, or at least an argument suggested by these views but perhaps not endorsed by Velleman, of direct concern to us may be formulated in this way. Necessarily, if one acts out of love, then one acts out of respect. If one acts out of respect, then one acts out of (Kantian) moral duty. It follows that if one acts out of love, one acts out of (Kantian) moral duty. Either of the following strategies may be adopted to assess the argument. On the first, we give detailed

analyses of the notions of Kantian respect, acting out of love, and acting out of Kantian moral duty and then revert to appraising the premises. On the second, assuming that we may acquire a sufficiently good grasp of these notions without detailed analyses, we explain and then evaluate the rationales for each premise. For the most part, we pursue the second strategy though, toward the end of this chapter, we exploit a toned down variation of the first.

8.5.1. An Outline of Velleman's Account of Love

A natural place to start is with a synopsis of Velleman's account of love. In "Love as a Moral Emotion," Velleman proposes that when we love someone, "we are responding to the value that he possesses by virtue of being a person or, as Kant would say, an instance of rational nature" (Velleman 1999, p. 365). Following Kant, Velleman claims that the value of a person is different in kind from the value of other things: "a person has a dignity, whereas other things have a price" (p. 364). The distinction between dignity and price corresponds to the distinction between ends that consist in possible results of action and ends that are "self-existent." The former, Velleman says, are objects of preference and choice and are comparative. The latter are not produced by action, and their value does not serve as grounds for comparing them with alternatives but as grounds for revering or respecting them as they already are. This value is incomparable in that "it calls for a response to the object [that has this value] in itself, not in comparison with others" (p. 364). Love then, is a response to (as a result of being aware of) the incomparable value possessed by a person in virtue of the person's being a self-existent end.

Velleman further proposes that love is an *arresting* awareness of such value. It is so in that, in responding to the incomparable value of a person in the manner constitutive of love, our defenses against being emotionally affected by the other are lifted (Velleman 1999, pp. 361, 366). Elaborating, Velleman explains that conceiving of love as a response to a person's rational nature may seem odd if 'rational nature' is taken to denote the intellect. However, rational nature he says, is not the intellect. Rather, it is a capacity of valuation: "a capacity to care about things in that reflective way which is distinctive of self-conscious creatures like us" (p. 366). We are invited to think of a person's rational nature as his core of reflective concern (pp. 366–67). What we respond to then in loving a person, is the value that the person has in virtue of being a person. This value "inheres" in the capacity persons have to appreciate the value of self-existent ends or, in other words, the capacity persons have for loving others. So according to Velleman, "what we respond to, in loving people, is their capacity to love: it's just another way of saying that what our hearts respond to is another heart" (p. 365). Since, in loving another, we respond to their capacity to love us, we suspend our emotional defenses against them:

> [L]ove for others is possible when we find in them a capacity for valuation like ours, which can be constrained by respect for ours, and which therefore makes our emotional defenses against them feel unnecessary. [Note omitted.] That's why our capacity for valuation, when facing instances of itself, feels able to respond in the manner constitutive of love, by suspending our emotional defenses. Love, like respect, is the heart's response to the realization that it is not alone. (Velleman 1999, p. 366)

We can now explain what Velleman sees as an intimate connection between love and Kantian duty. Velleman conceives of one's love for another as a response of the one to the qualities in the other in virtue of the having of which the other is a person. However Kantian respect is itself a response to what is common in its potential objects in virtue of which they are persons, namely, their rational nature or their capacity to appreciate the value of self-existent ends. Thus, Velleman regards the value to which we respond in loving a person as the same as that to which we respond in respecting them. Like love, Kantian respect is a response to the incomparable value one possesses by virtue of being a self-existent end. Velleman claims that he "regards respect and love as the required minimum and optional maximum responses to one and the same value" (Velleman 1999, p. 366). He suggests that the only fundamental difference between love and respect is one of degree. Both are responses of personhood in the one to personhood in the other; they are simply more or less intense responses to different degrees of intimate rapport.

If love though, just like respect, is a response to a person's capacity to appreciate the value of self-existent ends, a capacity that each person has in virtue of being a person, what explains love's selectivity or partiality? Why do we love only some but not all persons? In response, Velleman claims,

> Kant says that respect . . . for a person is a response to something that we know about him intellectually but with which we have no immediate acquaintance. According to my hypothesis, the value to which we respond in loving a person is the same as that to which we respond in respecting him—namely, the value of his rational nature, or personhood. But I have not said, nor am I inclined to say, that the immediate object of love is the purely intelligible aspect of the beloved. . . . The immediate object of love . . . is the manifest person, embodied in flesh and blood and accessible to the senses. The manifest person is the one against whom we have emotional defenses, and he must disarm them, if he can, with his manifest qualities. Grasping someone's personhood intellectually may be enough to make us respect him, but unless we actually see a person in the human being confronting us, we won't be moved to love; and we can see the person only by seeing him in or through his empirical persona. . . . One reason why we love some people

> rather than others is that we can see only into some of our observable fellow creatures. The human body and human behavior are imperfect expressions of personhood, and we are imperfect interpreters. Hence the value that makes someone eligible to be loved does not necessarily make him lovable in our eyes. Whether someone is lovable depends on how well his value as a person is expressed or symbolized for us by his empirical persona. . . . Another reason why we discriminate in love is that the value we do manage to see in some fellow creatures arrests our emotional defenses to them, and our resulting vulnerability exhausts the attention that we might have devoted to finding and appreciating the value in others. We are constitutionally limited in the number of people we can love; and we may have to stop short of our constitutional limits in order to enjoy the loving relationships that make for a good life. (Velleman 1999, pp. 371–72)

8.5.2. An Assessment of Velleman's Account

Velleman submits that "actions cannot genuinely be performed out of love without also being performed out of respect—and hence out of duty, though a joyous rather than grudging duty it is" (Velleman n.d., p. 21). Recall the argument with which we are concerned (again, we caution that Velleman may not subscribe to any such argument): Necessarily, if one acts out of love, then one acts out of respect. If one acts out of respect, then one acts out of (Kantian) moral duty. Hence, if one acts out of love, one acts out of (Kantian) moral duty. The rationale for the first premise rests on Velleman's view that both love and respect are responses to the same value that persons have in virtue of being persons. The rationale for the second appears to be that acting out of Kantian respect is sufficient for acting out of Kantian duty.

We may assess the rationale for the initial premise by asking whether the view that both love and respect are responses to the same value that persons have *qua* persons can indeed be squared with Velleman's view that whereas love is partial, respect is impartial. Focusing on the former, first, some property theories or accounts of love are especially vulnerable to what may be dubbed the "selectivity" or "partiality" problem. Derek Edyvane highlights the concern perspicuously:

> The extreme version of the properties view says something like the following; I love the person who possesses properties, x, y, and z, and I shall remain in love so long as he or she retains properties x, y and z. Should he or she lose these properties, the basis for loving will also be lost, taking with it any reason I might have had to remain committed. I think this view must be rejected for its failure to reflect our experience of love. We would certainly like to think, and very often it is the case, that love's bond is stronger and far less conditional than this kind of properties account would seem to imply. (Edyvane 2003, p. 62)

If love though, as Velleman theorizes, is a response to a property that *all* persons share roughly, the property of *being such that one has the capacity to value others as persons,* one would have thought that Velleman's account is particularly vulnerable to the selectivity problem. An essential aspect of Velleman's response to this worry, as we have seen, consists in his views that love disarms our emotional defenses, making us vulnerable to the other; and that not all persons are successful in arresting these defenses of ours. Consider, first though, love between parent and child. It seems highly dubious that when we love our children, we *suspend* our emotional defenses against them for the simple reason that our defenses against being emotionally susceptible to our children were never typically up in the first place.

Consider second, one's love for God. Velleman's thesis that love arrests our tendencies toward emotional self-protection from another agent seems to imply that, if Perry loves God, then God has disarmed Perry's emotional defenses against God. How though, *could* Perry have such emotional defenses against God? No person has effective defenses of any sort against God, and if he is a reasonable individual, Perry would be cognizant of this fact. How then, could God disarm emotional defenses of Perry against him? Reverse, now the order of love. Assume that God loves Perry. Then Perry disarms God's emotional defenses against Perry. But this result is, if anything, more bizarre than the former: why should an entity such as God have emotional *defenses* against any creature?

Imagine finally, a class of beings whose members are "emotionally transparent"; they have no tendencies toward others of emotional self-protection but are very much like us in other respects such as cognitive sophistication. They may bear the costs of being thus open but this is their fate. If love arrests our tendencies toward emotional self-protection from others, and these emotionally open creatures have no such defenses, then it would seem that they are incapable of love. This seems suspect.

Another component of Velleman's response to the selectivity problem appears to reside in his view that love involves valuing one's beloved as special and irreplaceable. Velleman submits that though what one values when one loves is a value that everyone has, it is a value with a "dignity" rather than a "price": a value to which it is inappropriate to respond by "comparing or equating one person with another" (Velleman 1999, p. 367). He underscores the point that when one judges that one's beloved has a value that everyone shares, this value calls for one to appreciate or value one's beloved as non-substitutable:

> [W]e can judge the person to be valuable in generic respects while also valuing her as irreplaceable. Valuing her as irreplaceable is a mode of appreciation in which we respond to her value with an unwillingness to replace her or to size her up against potential replacements. And refusing to compare or replace the person may be the appropriate response to a value that we attribute to her on grounds that apply to others as well.

> The same value may be attributable to many objects without necessarily warranting substitutions among them. (Velleman 1999, p. 368)

It is puzzling, though, why the fact, if it is indeed one, that love is a response to a value that calls for one to appreciate one's beloved as non-substitutable, should speak effectively to the selectivity conundrum. An analogy may help to reveal the problem. Frankfurt suggests that "The focus of love is not those general and hence repeatable characteristics that make his beloved *describable*. Rather, it is the specific particularity that makes his beloved *nameable*—something that is more mysterious than describability, and that is in any case manifestly impossible to define" (Frankfurt 1999, p. 170). It is almost as if that on Frankfurt's view, the focus of love is the beloved's haecceity or "thisness." Even on such a haecceity account, since each person has a haecceity if some person has a hacceity, one can surely intelligibly ask why the focus of one's love is *this* haecceity rather than another. Similarly, suppose each person, *qua* person, has the property of *being such that one's value—his or her value—cannot be compared or equated with that of any other person,* and that love is a response to such a property. One can still intelligibly ask why Perry responds to this property in Precious rather than to the property in Princess.

In sum, these considerations that tell against the effectiveness of Velleman's replies to the selectivity problem exert pressure against the view that love is a response to the same value as the value to which respect is a response. There are, in addition, anxieties from the other direction: respect, Velleman says, is impartial but it is not clear how this view about respect is to be sustained if love and respect are responses to the same value.

Velleman, as we have indicated, proposes that love as an emotional response, unlike respect, calls for a particular mode of contact or rapport between the lover and the beloved. The "loving response to a person requires us to enjoy an especially intimate rapport with him, and with personhood as instantiated in him" (Velleman n.d., p. 24). The role of the idiosyncratic features of the people we love, such as the way they walk, or talk, or look, Velleman says, is not that these features are objects or stimuli of love. Rather, they are "avenues" through which the necessary acquaintance or rapport is attained.

However, just as it is true that we do not love everyone, even assuming that love is a response to a generic value instantiated by each person, so it is true that we do not respect everyone, and this is true even if we have an intellectual grasp of the fact, again if it is one, that each person has the value to which respect is a response. Should one be partial to Velleman's view that respect is a response to the value of persons *qua* persons, this truth—that we respect some but not all people—strongly suggests that different people are drawn into rapport with others "along different avenues," the rapport at issue being that which is characteristic of respect. It may well be that one is drawn into the characteristic rapport with the other

because one admires the intelligence of the other, or the religious views of the other, or the political vision of the other, or the aesthetic sensibility of the other, or the sense of moral fairness of the other. Although, then, assuming that respect is a response that every person can have to a value possessed by each person, respect like love is, nevertheless, partial because different people are drawn into the relevant acquaintance with others to different degrees and along different avenues. In other words, Velleman's view that love and respect are responses of different intensities to the same value seems to imply that respect, just like love, is personal and partial because human beings can access the value to which respect is a response only via selective idiosyncratic routes. Not every person is open to others in a manner that would be required if, given Velleman's account of respect, respect is to be impartial.

Velleman's arresting thesis, or at least a thesis suggested by Velleman (though, perhaps, not endorsed by him), that actions cannot genuinely be performed out of love without also being performed out of duty is sustained by the intermediary premises that love and respect are responses to the same value that persons have in virtue of being persons, and that an action's being performed out of respect suffices for its being performed out of Kantian duty. We note that whether moral duty is to be identified with Kantian duty is controversial; it is something that requires defense.[8] Even, then, if one is partial to the view that love and respect are responses to the value that persons have as self-existent ends, there is room to disagree with the further contention that acting out of respect suffices for acting out of moral duty.

There is a final set of considerations against the view that one acts out of love only if one acts out of (Kantian) moral duty. Wholly intrinsically motivated actions exempted, in the previous chapter we proposed that when one acts from love, one is commendable, and that when one acts from moral duty, one is morally praiseworthy (assuming that freedom and other requirements of both these varieties of normative praiseworthiness are satisfied). Commendability seems to require action partly in light of the belief that one is doing what love demands; similarly, moral praiseworthiness presupposes action partly on the basis of the belief that one is doing right for right's sake. It is surely possible that, on some occasion, one may act in light of the belief that what one does is what love requires without, on that occasion, acting in light of the belief that what one does is what duty requires. And so it seems, it is surely possible that one can act out of love without acting out of duty. There is an alternative, similar route to the same conclusion. If one acts out of, for instance, justice, then one must satisfy an appropriate knowledge or awareness condition: one must believe that one's option is just, or more typically, one must believe that the option has the salient features of justice. Similarly, if one acts out of duty so that one is morally praiseworthy for what one does, one's act must issue at least partly from the belief that the act is morally right. Even

if, as some believe (Pettit and perhaps Frankfurt), that there is no analogous awareness condition for acting out of love, an agent can be moved to act out of duty with no thought whatsoever to love; love need not be any part of the motivational blend that gives rise to the action. It seems, consequently, that acting out of love may be entirely divorced from acting out of duty.

9 Love, Determinism, and Normative Education

9.1. INTRODUCTION: DETERMINISM, LOVE'S REQUIREMENTS, AND FOR WHAT SHOULD WE EDUCATE?

We have defended the thesis that the value for us of lovable behavior is essentially a function of our being commendable for such behavior against various objections. In this chapter, we first invoke this thesis to shed further doubt on Pereboom's proposal that hard determinism leaves intact relations of love. We then, in a very tentative fashion, address the issue of whether causal determinism poses a threat to the requirements of love and to appraisals of commendability and censurability. Finally, we conclude with a reexamination of the aims of education.

9.2. PEREBOOM'S VIEWS REVISITED

The central tenet of Pereboom's hard incompatibilism is that we do not have the freedom that moral responsibility requires. However, it is also an intriguing part of this position "that a conception of life without this sort of free will would not be devastating to our sense of meaning and purpose, and in certain respects it may even be beneficial" (Pereboom 2002, p. 477). As we recorded, the defense of this latter claim rests partly on the view that although hard incompatibilism undermines judgments of moral praise- and blameworthiness, this species of incompatibilism does not threaten morally deontic appraisals; presents no significant obstacles to achieving what makes our lives fulfilled, happy, satisfactory, or worthwhile; is consistent with an acceptable position on managing criminal behavior; and does not jeopardize important interpersonal relationships including relationships of love. We redirect attention to this last item. Although Pereboom has several interesting things to say about why hard incompatibilism does not endanger love or relationships of love, the following passage is especially noteworthy:

> Is it plausible that loving another requires that she be free in the sense required for moral responsibility? One might note that parents love

> their children rarely, if ever, because these children possess this sort of freedom, or because they freely (in this sense) choose the good, or because they deserve to be loved. Moreover, when adults love each other, it is also seldom, if at all, for these kinds of reasons. Explaining love is a complex enterprise. Besides moral character and action, factors such as one's relation to the other, her appearance, manner, intelligence, and her affinities with persons or events in one's history all might have a part. But suppose we assume that moral character and action are of paramount importance in producing and maintaining love. Even if there is an important aspect of love that is essentially a deserved response to moral character and action, it is unlikely that one's love would be undermined if one were to believe that these moral qualities do not come about through free and responsible choice. For moral character and action are lovable whether or not they merit praise. Love of another involves, most fundamentally, wishing well for the other, taking on many of the aims and desires of the other as one's own, and a desire to be together with the other. Hard incompatibilism threatens none of this. (Pereboom 2001, p. 202)

Lovable behavior, it is agreed, is frequently a vital component of loving relationships. We have seen that, according to Pereboom, if a person is morally praise- or blameworthy for a mental action, such as a decision, the production of the decision must be something over which the agent has control, and the agent is not morally responsible for the decision if sources over which she has no control ultimately produce it (Pereboom 2001, pp. 4, 47; 2002, p. 478). By Pereboom's view, this principle—*Principle O*—captures a requirement of ultimate origination for moral responsibility. Recall that Pereboom calls events for which factors beyond the agent's control determine their occurrence "alien-deterministic events"; he dubs events that are not produced by anything at all "truly random events"; and he designates the range of events between these two extremes—for which factors beyond the agent's control contribute to their production but do not determine them, while there is nothing that supplements the contribution of these factors to produce the events "partially random events." With respect to moral blameworthiness, Pereboom says that to "be blameworthy is to deserve blame just because one has chosen to do wrong. Hard incompatibilism rules out one's ever deserving blame just for choosing to act wrongly, for such choices are always alien-deterministic events, or truly random events, or partially random events" (Pereboom 2001, p. 140). Hard incompatibilism undermines moral praiseworthiness for similar reasons. We may summarize the relevant view in this manner: according to Pereboom, hard incompatibilism undercuts moral praise- and blameworthiness because hard incompatibilism precludes our ever being ultimate originators of any of our actions.

Now consider the pertinent view concerning censurability—blameworthiness from the point of view of love—that Pereboom would presumably

endorse: *To be censurable is to deserve blame from love's standpoint just because one has chosen to do what love forbids*. It would seem that if one accepts the view that hard incompatibilism rules out our ever deserving moral blame (or moral praise) just for choosing to act morally wrongly (or as we morally ought to) because we are never the ultimate originators of such choices, then one should equally accept the view that hard incompatibilism rules out our ever deserving blame (or praise) from love's standpoint just for choosing to act wrongly (or as we are obligated to), where these deontic assessments of wrong or obligation are assessments from the point of view of love. This is because again, we are never the ultimate originators of such choices. More succinctly, if there is a requirement of ultimate origination for moral praise- and blameworthiness then, in the absence of convincing reason to believe otherwise, there should be such a requirement for commendability and censurability as well. So, if hard determinism undermines moral praise- and blameworthiness, then it undermines commendability and censurability as well.

Let us take it then, that hard incompatibilism undermines the truth of judgments or ascriptions of normative responsibility from love's standpoint if it undermines the truth of such judgments or ascriptions from morality's standpoint. If what we value though, in loving behavior is essentially a function of being commendable for such behavior, contrary to Pereboom, hard incompatibilism *will* undermine relations (or some of their components) of love. The "lovable behavior" that remains intact in hard incompatibilist worlds is ersatz lovable behavior and not lovable behavior proper.

We have, however, argued that Pereboom's "combined generalization and counterexample strategy" that is called upon to establish the result that hard incompatibilism undermines the truth of judgments of moral responsibility is suspect. Hence, it stands to reason that this strategy cannot be invoked to show that hard incompatibilism threatens commendability or censurability. It is thus open to libertarians and compatibilists to advance and defend conditions under which a person is praise- or blameworthy from love's standpoint for her conduct.

9.3. DETERMINISM AND THE REQUIREMENTS OF LOVE

In Chapter 6 (Section 6.3) we adumbrated an argument for the view that determinism threatens the morally deontic judgments of right, wrong, and obligation. Briefly, the argument turns on two premises: first, the truth of these judgments presupposes freedom to do otherwise and, second, it is highly plausible that determinism effaces such freedom. Essential to the first premise are the "freedom-relevant principles" of obligation: the principle that "ought" implies "can" and the analogous principles concerning "right" and "wrong." If no morally deontic judgments are true

in a deterministic world, then determinism would also undermine moral praise- and moral blameworthiness *if* the following principles were true:

> Praise-1: An agent is morally praiseworthy for performing an action only if it is morally obligatory (or morally right) for the agent to perform the action.
>
> Blame-1: An agent is blameworthy for performing an action only if it is morally wrong for the agent to perform the action.

However, Praise-1 and Blame-1 are both highly controversial; indeed, as we have previously indicated, we believe that they are false. We have proposed elsewhere that one can be morally blameworthy for an action even if it is not wrong for the agent to perform the action (Haji 2002, pp. 162–96; 2001*b;* 1998, pp. 140–50) Should the agent act in light of the (non-culpable) but false belief that she is doing moral wrong, she may be morally blameworthy for what she takes to be a wrongdoing provided various other requirements of blameworthiness, such as that she acted freely, are satisfied. We do not intend in this work to defend fully our stance that Blame-1 and Praise-1 be jettisoned in favor of principles that require instead, that the agent act at least partly on the basis of the belief that she is doing moral wrong in order to be morally blameworthy, and that she act at least partly in light of the belief that she is doing moral right for right's sake to be morally praiseworthy (when these principles are taken to be competitors to Blame-1 and Praise-1). We did, though, say some pertinent things in support of our view in the preceding chapter.[1] We simply note that even if these weaker "belief replacements" of Blame-1 and Praise-1 were acceptable, judgments of moral praise- and blameworthiness in a deterministic world would be irrational in the sense that they would all be predicated on morally deontic beliefs that are false.

It is interesting to inquire whether determinism similarly threatens love's prohibitions or requirements—what we may abridge somewhat cumbersomely as "love's deontic prescriptions" or simply as "L-prescriptions"—and whether ascriptions of censurability and commendability are similarly irrational. Concerning the former, if an argument analogous to the one that we have outlined for the conclusion that no morally deontic judgments are true in a deterministic world is to be put to service to show that no deontic prescriptions from love's standpoint survive in a deterministic world, we would need to establish that if something is obligatory from the point of view of love for an agent, then the agent can do that thing; as we shall say, we would need to sustain the view that "ought-L" implies "can." (Similarly, we would be required to substantiate the principle that "wrong-L" implies "can" as well as the principle that "right-L" implies "can.")

At least two standard tactics can be pursued in tandem to defend the principle that the "ought" of morality—"ought-M"—implies "can": rebut

arguments against, and marshal "theory-based" support for, the principle; regarding the latter, demonstrate, for instance, that a highly encouraging analysis of the concept of moral obligation includes the principle as a theorem. As we did before, refer to the principle that "ought-M" implies "can" as "**OMC**"; and dub the principle that "ought-L" implies "can," "**OLC**." It should be evident that neither of these tactics is promising when it is **OLC** that is at issue. Perhaps we could make some progress if there were grounds to believe that certain arguments against **OMC** that are controversial are less, and preferably, far less so, if recast as arguments against **OLC**. Here, we dwell on what we believe is one such argument.

Some people have attempted to reject **OMC** by appealing to genuine moral dilemmas, that is, by appealing to the possibility of genuine conflicts of all in or overall moral obligation. The argument is as follows. Assume that there are genuine moral dilemmas. Then there can be situations in which (ignoring temporal indices) an agent, *S*, ought to do *A* and ought to do *B* but cannot do both *A* and *B;* specifically, the agent cannot do the conjunctive act (*A* and *B*). As *S* ought to do *A* and *S* ought to do *B*, the agglomeration principle:

AGP: If O(*A*) and O(*B*), then O(*A* and *B*),

implies that *S* ought to do the conjunctive act (*A* and *B*). If **OMC** is true, then *S* can do this conjunctive act. But *S* cannot, in *S*'s dilemmatic situation, do this conjunctive act. Hence, it is concluded that **OMC** is false.[2] Some have questioned the agglomeration principle to escape the argument.[3] But even if one accepts this principle, the major premise of this line of reasoning against **OMC**—that it is possible for there to be genuine moral dilemmas—is highly contentious.[4]

Recast the argument as an argument against **OLC**—the "ought-from-the-point-of-view-of-love" implies "can" principle. Accept the agglomeration principle (now understood as a principle governing "ought-L" appraisals) and focus on the premise that there are basic "ought-L" dilemmas; love's requirements can genuinely conflict. We suggest that this premise is *less* controversial than the analogous premise involving moral obligation. Even what we regard as the most plausible consideration in favor of the possibility of basic moral dilemmas, the consideration that it is obvious that they can occur, is subject to skepticism. Zimmerman speaks directly to this concern:

> [I]t may be that some proponents of dilemmas believe that it is just obvious that basic dilemmas can occur. Cases such as Sartre's, that of Agamemnon at Aulis . . . , that of Sophie's choice . . . , and others are often presented as being clearly dilemmatic. Well, it is clear that a conflict of some sort is at issue in these cases, but is it clear that what's at issue is a basic dilemma? How could this be, even from the proponent's

> point of view? For they of course grant that there can be conflicts that are morally resolvable, and some of these may be very hard cases (in the sense that it is very difficult to figure out just what the solution is). If so, it would seem that it can never be obvious that a particular conflict constitutes a basic dilemma rather than merely a very hard but resolvable case. (M. J. Zimmerman 1996, p. 220; notes omitted)

However, now consider the claim that it is just obvious that love's requirements can genuinely conflict in that there can be cases in which some agent ought-L to do something and ought-L to do something else, but cannot do both. The claim appears to be borne out by certain cases involving symmetry. If an unfortunate mother must decide which of her two twins to save from certain death on pain of losing both, it *seems* compelling that she ought-L to save one, and that she ought-L to save the other, even though she cannot save both. Though we admit that we may well be mistaken on love's verdicts in scenarios of this sort, the consideration that love's requirements are particular and that they are relatively more agent focused than act focused, motivates our view. One way in which love is particular is that, as Frankfurt says, the bonds of love are not transferable (Frankfurt 1999, p. 169); and one way is which love is "agent focused" is that the lover takes on the interests of the beloved as her own. Frankfurt explains:

> Lovers are not merely concerned for the interests of their beloveds. In a sense that I shall not attempt to define but that I suppose is sufficiently familiar and intelligible, they *identify* those interests as their own. Self-love, in which the interests of the lover and the beloved are *literally* identical, is an unequivocally robust paradigm of this. As I emphasized . . . , the interest of the lover in his beloved is not generic. He does not love his beloved because to do so fulfills certain independently specifiable conditions that qualify it as a member of a certain class. If that were so, then his love would be satisfied by any other objects that might also belong to that class. In fact, however, love of a beloved object cannot be satisfied by anything except that very object itself. . . . The bond between a lover and his beloved is not transferable. A person cannot coherently accept a substitute for his beloved, even if he is certain that he would find himself loving the substitute just as much as he loves the beloved that it replaces. (Frankfurt 1999, pp. 168–69)

The requirements of love must heed love's particularity. Love requires, for example, that (generally) one save one's beloved rather than the stranger; morality may dictate otherwise. In addition, the requirements of love necessitate devoting special attention to the cares, concerns, or interests of the beloved rather than to those of some other person even though this may be contrary to what morality prescribes. In this way, we shall say that love is relatively agent focused. Revert to the dire predicament of the mother.

Moved by love, she is evenly devoted to both her children, assuming the cares and concerns of either as her own; she is equally "invested" in both. It seems that love *requires* such investment. A tragedy of love, then, it appears is that no matter which child she saves, she cannot avoid wrongdoing from the point of view of love. Such is love's devotional commitment.[5]

If there can be genuine conflicts of obligation of love, then perhaps **OLC** ought to be rejected. *If* this principle is suspect, *one path* of reasoning from the truth of determinism to the conclusions that love's deontic prescriptions and the appraisals of commendability and censurability cannot survive in a deterministic world is a non-sequitur.

9.4. EDUCATIONAL AIMS REVISITED

We now revert to a challenging question that we have delayed answering. If an overarching aim of education is to make certain that our children mature into normative agents, what kind of normative agency should we aim for and on what basis is such a decision to be made (Chapter 7, Section 7.2.1)? Before responding, we first deflect what appears to be an important skeptical challenge to the view that education *has* overarching aims, and we then briefly discuss proposed, prominent conceptions of what these aims are supposed to be.

9.4.1. Skepticism About the Aims of Education

Some people distinguish between "ideals" of education and "aims" of education, although these two terms are often used interchangeably. According to Doret De Ruyter, ideals are "things that people consider to be excellent, the optimum or the best" which they have not yet realized (De Ruyter 2003, p. 468). Sometimes she puts the point slightly differently. She says that ideals are images of excellences (pp. 468, 478). De Ruyter explains that if a person believes that something is an ideal, the person will be motivated to attain it, and she will be so motivated, in part, because she takes the ideal to be valuable and to be something that is goal-setting (pp. 471–73). De Ruyter differentiates two broad classes of ideals: the class whose members specify situations agents take to be excellent ("ideal situations") and the class whose members specify traits of character agents deem excellent ("character ideals") (pp. 469–70). At times, De Ruyter suggests that ideals *are* situations, at other times she takes them *to be* qualities (or "excellences") of character or images of excellences, at yet other times she thinks that ideals *are* virtues or values. Perhaps her view is that all these things, or pertinent states of affairs that have these things as constituents, are ideals. It is clear though, that she believes that a person may have a variety of ideals including moral, religious, social, political, economical, and aesthetic ones, and she submits that the "excellences of character" that are ideals

include "virtues such as courage, temperance, wisdom, justice, honesty and generosity" (p. 470). De Ruyter calls attention to another feature of ideals. She says that ideals have a personal nature in that people may differ on what they take to be ideals (pp. 472–73), and that something can be an ideal for one person but not for another because this other person, for instance, has already "achieved it" (p. 473). De Ruyter also emphasizes that "ideals are a sub-class of values" (p. 473). Owing to their status as "supreme values," De Ruyter claims that ideals should be part of one's conception of the good life. Ideals are "existentially important" because they "give direction, inspiration and incentive to make something special of one's life or to lead a flourishing life" (p. 475). She claims that one's ideals can influence one's conception of the good life and that the converse is also true. Further, De Ruyter proposes that the ideals that are moral excellences are crucially important "for the formation and composition of . . . personal identity" (De Ruyter and Conroy 2002, p. 509). She concludes that because of the "the existential importance of ideals as well as their stimulating force," ideals "are important in education and, therefore, that parents and teachers should offer ideals to children." (De Ruyter 2003, p. 476)

How are ideals to be distinguished from aims? De Ruyter suggests that this distinction, if it is a real distinction, is closely aligned with the distinction between what she labels "*educational ideals*" and "*ideals in education*": "The first refers to the ideal aims and practices of education, and the second to ideals that educators offer to children" (De Ruyter 2003, p. 476). Regarding the latter, De Ruyter explains that educators should present children with moral ideals, including the moral excellences (the germane virtues), and that they should steer children away from "immoral ideals" such as racist or sexist ones. When exposing children to these ideals, De Ruyter and Conroy (2002) recommend that an educator's own behavior should not be in discord with them. As for the former, De Ruyter ventures that educational ideals appear to be a sub-class of *educational aims*. Unfortunately, this characterization or mark of educational aims fails to provide us with the relevant distinction: we are still left in the dark about just what it is that differentiates aims from ideals. Perhaps De Ruyter's recommendation is that the distinction between ideals and aims is a distinction without a difference.

Let us start over. Unless one has special reason to believe otherwise, it would be prima facie highly implausible to deny that education has aims. Education, after all, appears to be goal directed (or as some say, "teleological"). In addition, when it comes to educating our children, we believe that some goals are worthy of pursuit, others not, or some more worthy of pursuit than others. Education is thus, also "value directed" or normative. We may distinguish between relatively specific educational goals or aims and relatively more general aims. As examples of the former, we think that children should be equipped with reading, writing, and effective oratory skills. As examples of the latter, we believe that ensuring that our

children develop into autonomous critical thinkers and morally responsible agents is a good thing. Regarding responsible agency, most of us assume the task, either consciously or not so consciously, of shaping our children into responsible agents. Fisher and Ravizza have some informative things to say about this sort of "moral training":

> When a child goes through the long, complex, and difficult process of "moral education," one might say that the child is becoming a "moral agent." Part of what it is to be a moral agent is to be a participant in the configuration of practices constitutive of moral responsibility. . . . [I]t will be helpful to consider aspects of a child's moral education in the "typical" case. . . . Even before children are fully responsible for their actions, we often find ourselves taking certain attitudes toward them that are in many respects similar to the full-blown attitudes of indignation and resentment (which are of course only appropriately applicable to morally responsible agents). . . . By adopting certain attitudes toward the child (and expressing them suitably)—by acting *as if* the child were a fully developed moral person—we begin to teach the child what it means to be such a person. Of course, this sort of training, with its characteristic set of parental attitudes and responses, is a central feature of the moral education of children. . . . But how exactly does this education "work"? . . . Parental responses to a child's behavior, as part of the typical process of moral education, seek to induce the child to accept a certain view of himself as an agent. The relevant notion of "agency" is a rather minimal notion, according to which the child sees himself as the source . . . of certain upshots in the external world. The sense in which the child sees himself as the "source" of these upshots is that he sees that their occurrence is caused—in a characteristic way—by *him*. The child is brought to see that his desires, beliefs, and intentions result in actions and upshots in the world. . . . Further, the child is typically invited to see that, when he exercises his agency in certain contexts, he can fairly be praised or blamed for his behavior. . . . Once a child has acquired this sort of view of himself, he can at least provisionally be *held* morally responsible for his behavior. . . . At this stage in the development of a fully morally responsible agent, the child is (at least provisionally) rationally accessible to the reactive attitudes. When we adopt such attitudes toward the child, we expect that they will be met with an appropriate response, and that the child will adopt an internal attitude toward himself that corresponds to the external attitude we adopt toward him. (Fischer and Ravizza, 1998, pp. 208–09)

We may identify the relatively general aims of education, such as ensuring that the child becomes an autonomous critical thinker and a morally responsible agent, as "overarching." This is simply because although relatively specific educational aims may vary, and sometimes considerably so

depending upon various contingences, such as economic or cultural ones, the overarching aims seem more basic, stable, and universal; they are (typically) a fundamental ingredient in a life that is good in itself for a child. The spectrum of educational aims, with highly specific ones at one extreme and highly general ones at the other, suggests that the distinction between ideals of education and aims of education is not so hard and fast. It would not strike us as out of the ordinary if it were proposed that one goal of education should be to foster the moral virtues. De Ruyter, as we have seen, takes such virtues to be ideals.

We should nip one thorn in the bud. Some skeptics maintain that the pursuit of educational ideals leads to frustration rather than to satisfaction in life.[6] Others claim that the pursuit of such ideals gives rise to fanaticism and indoctrination. Still others argue that the pursuit of ideals belies pursuit of "unattainable perfectionism"; what we should strive for is "anti-perfectionist realism." Whatever precisely the advocates of these views might mean by "ideals," their conceptualization of ideals should not be conflated with what we have referred to as the "overarching aims of education." It would be preposterous to claim, for instance, that training our children to think critically has anything essential to do with fanaticism.

This brief discussion on the aims of education raises two difficult challenges. One centers on the comparative significance of the aims. What, exactly, does it mean to say that some aims are more important than others? What is the precise import of this claim? The second, once again, concerns ideals of education. At least on the face of it, the challenge, this time, is directed against the very idea that education has bona fide aims in the sense of "aims" that we have advanced. We first direct attention to this challenge. We then address the significance of educational aims.

Mundane astuteness counsels that educating children is goal-bound, aimed toward attaining ends deemed worthy, and, is therefore, also value-based. Yet it may appear that Paul Standish has serious doubts about this goal-directed and value-oriented structure of education:

> Sometimes the question [of the aims of education] has been seen as the issue *par excellence* for philosophy of education. There is considerable merit to this view but also dangers of portentousness and pomposity. The preoccupation with aims may stand in the way of the more patient characterization of good educational practices that is of real benefit to practitioners. It tends to predicate the consideration of education on a teleological metaphysics, harboring a fallacy of essentialism. (Standish 2003, p. 223)

Although Standish admits that "scepticism about the giving of aims may seem like a kind of political irresponsibility" (Standish 1999, p. 41), he nevertheless seemingly attacks the very idea that education has aims. First, Standish claims that the assumption that there must be aims accords

with the "presumption in favor of rational planning" in the modern world (Standish 1999, p. 41). Rational planning is then interpreted as symptomatic of the reign of instrumental reason which is burdened with all the "sins of modernity," such as scientism, technicism, objectification, mechanical effectiveness, efficiency, performativity, technology, managerialism, and quality control. Standish proposes that at a more grammatical level, furthermore,

> it is worth instancing examples of valued practice where the aims are inexplicit or where there are no aims—or perhaps where talk of aims seems inappropriate. Indeed, some of the most important aspects of people's lives—their intimate relationships, for example—seem to be characterised in this way. Within such practices, there may be a great many smaller-scale practices in which aims can more or less be identified. But these are likely to be understood in the light of something which cannot be formulated in any tidy way and which would be inappropriately thought of in terms of aims. To ask for the aims of education may be like asking for the aims of a town. What, for example, are the aims of Aberdeen? The grammatical oddness here suggests that there may not be much sense in the question. The critic will respond that there are indeed aims of Aberdeen and these have been made quite explicitly by the members of the town's council, who have worked earnestly to devise their mission statement. A mission statement of this sort may or may not be desirable but it is clear that, although this may be an appropriate expression of the political intentions of a dominant faction, this hardly warrants their attribution to the town! While a town incorporates a diverse range of purposeful practices, it is not clear that aims of an over-arching kind can be given. The multiple smaller-scale projects which go to make up the life of the town will include in their number those where things do need to be planned out, sometimes systematically. But these will have their sense in the light of that larger purposiveness. . . . But if such statements of aims are indeed ungrammatical or prejudicial, this may be an unwarranted security, one which is apt to distort our practices. . . . By analogy, the suspicion which emerges is that stating the aims of education may lead to a kind of stifling. A seemingly logical progression leads towards systems of aims and objectives and to a preoccupation with performativity which dominates the curriculum. (Standish 1999, pp. 41–42)

To Standish's credit, this interesting passage reveals an important ambiguity in the concept of *being an educational aim*. Just as there is something grammatically suspicious with asking for the aims of a town, so one may believe there is grammatical oddness with inquiring about the aims of a life, or the aims of education. Citing Standish, "Persons, parents, and

teachers have aims, Dewey reminds us, not an abstract idea like education" (Standish 1999, p. 42). We concede that one aim of *educators* is to impart certain basic knowledge to our children. We can then say that, derivatively, education has such "low-level" or specific aims. It is, though, not grammatically untoward to inquire into what makes a (good) town good, or what makes a life good in itself for the person who lives it, or what makes (good) education good. (Needless to say, when attempting to respond to these questions, the sense of "good" would require clarification.) We previously documented that it is generally acknowledged that education is goal directed and value oriented. What we have characterized as education's relatively specific aims perspicuously expose education's goal-directed element, and what we referred to as education's relatively general aims highlight education's value-oriented facet. Our view, to be developed below, is that securing the "aims" of developing into an agent who is, for instance, an apt candidate for ascriptions of moral responsibility and who is a suitable candidate for appraisals of responsibility from the point of view of love, are (characteristically) vital constituents in the good life for the child. The child may be led to see himself as a fair target of the reactive attitudes. To facilitate such growth, the pertinent regimen of "training" is, in a clear manner, teleological.

There is another strand of Standish's skepticism concerning the goals of education that run through many passages in his germane works:

> The idea that . . . there should be some kind of categorization of aims, perhaps the better to identify the ones that best suit our circumstances—seems to distort what is at issue here. For it gives the impression that aims are things that might be chosen and then attached to means adopted or developed in order to realize them. To many, the good sense of such a procedure will seem as clear as the light of day. But what goes wrong here has to do with a failure to understand the extent to which aims are internally related to certain kinds of practice. (Standish 2003, p. 222)

Remarking on some of John Dewey's observations on the vice of externally imposed ends—ends, for instance, dictated to teachers from superior authorities who, in turn, accept them from what is current in the community—Standish says,

> In contrast to the above [that is, to externally imposed ends], aims are to be understood first in terms of the purposiveness of human activity, as internally related to particular activities. Truly general aims, if such there are to be, should broaden the outlook, enabling a wider and more flexible observation of means and exposing the endless connections of particular activities: teaching and learning should lead indefinitely into other things. (Standish 1999, p. 42)

Echoing these views of Dewey, Standish submits,

> If an aim is an external end to which the means is related only instrumentally, then education in liberal terms is indeed aimless; in *The Sovereignty of Good* Murdoch speaks of virtue as pointless. But clearly this is not the only possibility and it should not stop argument. Modern philosophies of liberal education have recognized correctly that the aims of education must be seen in terms of the good. (Standish 1999, p. 48)

Standish contrasts education that unfolds in accord with "externally imposed ends" with an alternative that he finds in, among others, Plato's, Dewey's, and especially Iris Murdoch's writings, a *via negativa* that seeks to locate "the good" in opaque, oblique, tentative and evocative ways (Standish 1999, p. 48). Summarizing aspects of this alternative, Standish writes,

> A literarily crafted philosophy of education would open the possibility of a way of thinking which would unsteady the discourse of liberal education. It would do this not to jettison liberal education but to resist the limitations which its monologism makes it subject. In doing so it would keep liberal education open to that ancient sense of the good which modern formalistic and naturalistic tendencies have subdued or obscured. Sceptical of the direct representation of the good it would locate itself in a recollection of what has been said before, in a response to texts going beyond anything which could be made fully present. Its withholding and humility, sometimes its renunciation of the claim to know, would themselves be characteristics of that intimation of the good which defies clear statement in a set of aims. This is the kind of thing in which teacher and learner might well be enthralled. (Standish 1999, p. 48)

The passages we have cited seem to confirm that Standish does *not,* in the end, renounce the view that education has aims. Rather, one of Standish's insights appears to be that the overarching ends of education are to be characterized in terms of the good. Reconsider some of the candidates we have proposed as apt candidates for education's overarching ends: turning our children into morally or normatively responsible agents and striving to ensure that our children are autonomous critical thinkers. We dwell on two features of these "aims." First, they are not or at least they do not seem to be "things that might be chosen and then attached to means adopted or developed in order to realize them" (Standish 2003, p. 222); they are not "externally" or "extrinsically" related to certain kinds of practice.[7] These aims are not, in these respects, relevantly analogous to the aim, for example, of ensuring that our children turn into the best computer programmers. Aims of this sort may well be consciously adopted, educators being instructed to find or develop the best means to achieve them. In contrast, turning children into morally responsible agents is a *precondition*

for "behaving purposefully and meaningfully in important aspects of life" (p. 222). Our interpersonal relationships or at least central species of such relationships, for instance, take it for granted that we are morally responsible agents. We need to understand the vocabulary of moral discourse and attendant moral practices to see ourselves and others as agents of a certain sort who can participate effectively in social relationships. Becoming a moral agent is a precondition of "admittance" into a world in which significant behavior becomes intelligible and meaningful only if it is viewed or interpreted through the lens of moral categories.

Second, the overarching aims that we have identified are plausibly thought of as being vital elements in a life that is intrinsically good for the one who lives it (or so we shall argue). It is in this way that these overarching aims are characterized in terms of the good. In addition, the good life for a person is, presumably, the life the person should seek. So there is a non-trivial sense in which the overarching aims are teleological. It is noteworthy that Standish himself cannot fully escape from the "teleological metaphysics" of education and the "essentialism" of educational aims in his endeavors to characterize *good* educational practice. His attempts at illumination invoke the idea of a *via negativa;* but a negative way to something valuable is, surely, still a *way* to that end. Further, it appears that Standish's criticism that the educational practices in our industrial and globalized economy are "limited and debased" appears to acquire legitimacy only against his vision that education *does* have certain bona fide overarching aims which are to be understood in terms of the purposiveness of human activity and as "internally related" to particular practices (1999, p. 42). We may conclude that the thesis that (moral) education is goal-directed and value-oriented is a thesis that should be given serious consideration.

9.4.2. A Duality of Aims: Liberal and Non-Liberal Education

A central theme in Aristotle's early sections of the *Nicomachean Ethics* (Book I) is that the good life is the life of flourishing and virtue. To achieve a state of well-being, proper social institutions are necessary. The political setting must enable people to cultivate the peculiarly human excellences, the virtues, which are necessary for the good life. Indeed, Aristotle proposes that the state should actively encourage people to inculcate the virtues in order for its citizens to flourish. Their flourishing, in turn, will ensure that the polis itself flourishes as well. For this reason, ethics is inextricably associated with the political order. If we have this sort of picture of human welfare in mind, then it should come as no surprise that moral and political philosophy ultimately inform conceptions of the overarching aims of education.

Although this vision of the good life has been contested, the suggestion that morality and political philosophy fundamentally bear on conceptions of the overarching ends of education has had a lasting influence.

For example, witness the "liberal" versus "non-liberal" paradigms of education (Noddings and Slote 2003; Callan and White 2003) The first incorporates Kantian moral philosophy, Rawlsian political philosophy, and Kohlbergian developmental psychology; the second virtue ethics, communitarianism, and an ethic of care. As John White observes, the debate between liberals and non-liberals is far more than a theoretical diversion for philosophers, political scientists, psychologists, and educational theorists: "At stake are rival understandings of what makes human lives and the societies in which they unfold both good and just, and derivatively, competing conceptions of the education needed for individual and social betterment" (Callan and White 2003, p. 96). If we take this view to heart, the overarching ends of education must speak to both individual welfare and social betterment. We shall focus primarily on individual welfare.

What conceptions of the ends of education do we find in the liberal and non-liberal paradigms? We may take our cue from what Nel Noddings and Michael Slote propose concerning moral education:

> There seem to be three main philosophical theories of morality (or four, if we separate virtue ethics and communitarianism) that could potentially influence current understanding of moral education. Virtue ethics and mainstream communitarianism would naturally encourage a form of moral education in which schools and parents would seek to inculcate good character in the form of specific (labeled) habitual virtues. Kantian/ Rawlsian rationalism/liberalism would seemingly encourage moral education to take the form of developing certain capacities for moral reasoning and certain very general principles [derived from a general duty of respect for the autonomy and dignity of every person] that can be applied to different moral dilemmas or decisions. Finally, an ethic of care would most naturally see moral education as a matter of children's coming to an intelligent emotional understanding of the good or harmful effects of their actions on the lives of other people as well as deepening understanding of defensible ways to live their own lives. Care involves caring for oneself as well as others. (Noddings and Slote 2003, p. 349)

Simplifying somewhat, Noddings and Slote suggest that the aims central to the liberal paradigm are tied to personal autonomy, moral reasoning, and critical thinking, whereas those at the heart of the non-liberal paradigm are affiliated with good character, moral sentiment, and caring relationships involving benevolence and kindness. Aims of the first type are concerned with encouraging self-conscious and conscientious attention to one's own goals, values and choices, and promoting obedience to universal (moral) rules and principles. Aims of the second sort are associated with enforcing spontaneous other-directed reactive attitudes and feelings, and inculcating particular acts of caring.

White, who builds on the liberal position R. S. Peters, P. H. Hirst, and R. F. Dearden enunciated in the 1960s, depicts liberal education and its goal as follows:

> Education aims at promoting pupils' personal well-being. In a liberal-democratic society, . . . this will include personal autonomy. (White 1990, p. 36)

> Autonomy depends on the existence of options. Education cannot supply these, but it can make students aware of them. Its job is partly to open up horizons on different conceptions of how one should live—ways of life, forms of relationship, vocational and nonvocational activities. But a broad understanding of options is not enough. Autonomous agents also need to understand themselves. They need to interpret their major goals and establish priorities among them, and to discern possible psychological obstacles arising to their self-directness. . . . They need also to be equipped with qualities of character. They have, for instance, to be able to withstand pressures to conform to what authority or public opinion want them to do. For this they require the critical independence of thought to asses others' arguments, as well as the moral courage to stand up for their own views. Exercise from as early an age as practicable in making choices and reflecting on these is a further requirement—as is a whole-heartedness of commitment to activities of their own choosing. (Callan and White 2003, p. 97)[8]

In contrast to emphasizing values of individual choice and personal well-being, non-liberal educational theorists call attention to community values, traditions, and good habits, and they underscore the desirability to care for other people and to act out of immediate concern for the welfare of others.

We have claimed that striving to ensure that our children develop into morally responsible agents, into agents who are appraisable from the point of view of love (or, in short, into loving agents), and into autonomous critical thinkers are primary aims of education. This conception of (some) of education's overarching aims may give the initial impression that we are more in sympathy with the liberal than with the non-liberal camp. However, this would be premature. Our position, to be developed below, distils to the following. Essential elements of education make for human welfare or flourishing. When thinking about the overarching aims of education, the following question (among others) should guide us: What do we offer to the child that makes the child's life good in itself for her? Our proposal is that the three elements we have identified—being morally responsible, being commendable or censurable, and being autonomous critical thinkers—are (typically) vital elements in the good life.

9.5. EDUCATIONAL AIMS AND THE GOOD LIFE

Let us backtrack for a bit. We may act from duty, love, self-interest, religious conviction, aesthetic inspiration, and other things. Conditional upon the normative stance or standpoint at issue, we may, thus, be morally praise- or blameworthy, praise- or blameworthy from the point of view of love, and so on. Whether we habitually act out of, for instance, moral duty or love depends largely on how we are raised. If we are brought up to heed morality, we will be generally disposed to act from moral duty; if brought up to heed the dictates of love, we will be generally disposed to act from love. How, then, should we be relevantly raised? What, precisely, is in question requires clarification.

In asking how we ought to be relevantly raised or how our children ought to be relevantly raised, one may be asking how we ought morally, or, say, from the point of view of love, to raise our children. The "ought" under consideration takes on the sense of a specific variety of obligation. If it is the moral "ought" that is of concern, one possible answer is that we ought morally to raise children so that they heed a variety of norms, not just moral ones. This is not, however, the question we have in mind.

Alternatively, acknowledging that there are varieties of normative responsibility, each associated with its own type of normative agency, we can raise children into different sorts of normative agent. Which sort or sorts are to be favored? One may start by asking which variety of normative responsibility is most significant or important. In turn, this assessment depends upon which evaluative standpoint or, if we want, which standard of "obligation"—moral, prudential, legal, etiquettical, that of love, and so on—rightly affiliated with the variant of normative responsibility under consideration, is itself relevantly most important. The proposition is that we raise or educate our children with an eye toward ensuring that they turn into the sort or sorts of normative agent deemed most important in this sense. Exactly what notion of importance though, is at issue?

One suggestion is that the most important normative standard is "overriding" roughly in the sense that, of all the normative "oughts," its "ought" takes precedence: when its "ought" requirements conflict with the requirements of other "oughts," its "ought" requirements are most weighty. For instance, suppose that moral obligation is overriding. Then if, as of some time, t, a person ought morally to do action A at t^* (where t^* may be identical to or later than t), and ought (say legally) to do B at t^*, but cannot then do both, then she plain ought to do A at t^*. The phrase, 'she plain ought to do A at t^*,' is meant to capture the idea that moral "oughts" are more "normatively significant" or override other "oughts" like legal ones.[9] There are though, severe difficulties with this view. First, it is not clear that there is anything like "plain ought," an overarching standard that passes impartial and final judgment on the relative normative stringency of specific "oughts" such as moral, legal, and prudential "oughts."[10] Second, even

if there is such a standard, it is not obvious what verdict this standard delivers. For instance, it is not evident that the standard of "plain ought" rules that moral obligation is supreme (Haji 2002, pp. 221–44). Third, assuming there is an overarching standard and assuming it delivers some concrete verdict, for example, the ruling that moral obligation is overriding, it is still an open question whether our lives would go best for us if we conscientiously strove to do what is morally obligatory or right and conscientiously strove to avoid doing moral wrong. Perhaps our lives would go better if we were always to act from love. Finally, there is the concern of which standpoint or standpoints agents in fact *take* to be important in guiding their conduct independently of which standpoint, if any, is overriding. Regarding this concern, it is sensible to suppose that no single standpoint—not even the moral standpoint—enjoys a privileged status in the lives of most people. Reflecting on our day-to-day dealings with others, it seems that an agent in many or most situations is not wedded to any one evaluative standpoint. As we have previously acknowledged, one does not usually commit oneself to acting on (or despite) for instance, prudential considerations or those deriving from the heavy hand of tradition no matter what the situation in which one finds oneself. It appears, rather that, in ordinary life, many of us reveal a disposition to be relatively flexible with our values, taking moral considerations to be more important in some situations than, say, legal or prudential or aesthetic ones, but reversing our commitments in different situations, and in yet others paying no heed to morality at all but acting out of or despite love or friendship. Thus, whereas Jenny the artist may take caring for helpless children to trump a commitment to her artistic enterprises, she may not take some other equally compelling moral consideration—such as a contributing as a nurse's aid to a crisis center after some natural catastrophe—to do so.

Our preliminary suggestion is that in the end, the choice regarding which normative standard is most significant and, hence, which sort of normative agency we should aim for in educating our children be pragmatic in this sense: to the extent that this is possible, let life or experience be our guide. An undoubtedly controversial suggestion we advance is that love is of paramount significance. Our lives would be far intrinsically better for us if love and care were emphasized in our dealings with others. If this is plausible, then the sort of normative agency associated with love should be of singular importance in the normative education of our children.

We now develop this suggestion. Perhaps the rock bottom aim of education is to raise our children in such a fashion that, of the many different ways in which their lives could turn out, each child gets a life that is good in itself—that is intrinsically valuable—for him or her. The manner, then, in which we should proceed in educating our children to fulfill this basic aim, as Aristotle seems to have recognized, cannot be ascertained without treading into the deep waters of axiology. The key question that needs to be addressed is this: what makes a life good in itself for the one who lives

it? This question should not be confused with another: what sort of life contributes to the intrinsic value of some *world?* A life that is highly valuable in itself for the one who lives it may have very little to do with the overall intrinsic value of a world; it may have low extrinsic value. It is, thus, important to distinguish between educating with an eye toward making the life that the child lives better, in itself, for the child and educating with an eye toward making some world intrinsically better. Our concern (in this work) is with the former and not with the latter question. We do not deny that the latter question should be of paramount concern in the philosophy of education. In what follows, however, we work from the assumption that we should educate for personal well-being, leaving it open whether this aim is only one among others of education's basic aims. Another question of pressing urgency for our interests is the following: what role, if any, does acting from love play in contributing to the intrinsic value of our lives? A similar question can be raised in connection with the aims of striving to ensure that our children develop into morally responsible agents and into autonomous critical thinkers.

9.5.1. Intrinsic Value and Attitudinal Hedonism

We need minimally to get clear on when a person's life is good in itself for that person—in what does personal welfare consist? Any serious discussion in the philosophy of education of the justification of overarching educational aims is inherently associated with the debate concerning alternative life-ranking axiologies. Regarding well-being, Derek Parfit distinguishes a number of theories:

> On *Hedonistic Theories,* what would be best for someone is what would make his life happiest. On *Desire-Fulfilment Theories,* what would be best for someone is what, throughout his life, would best fulfil his desires. On *Objective List Theories,* certain things are good or bad for us, whether or not we want to have the good things, or to avoid the bad things. (Parfit, 1984, p. 493)

It should go without saying that we cannot in this work undertake the enormous burden of doing justice to the issue of what makes a life intrinsically good for the one who lives it. Our aspirations, in this connection, are very modest. We intend to sketch what we believe is a promising program of investigation.

Every axiology—roughly, every theory of intrinsic value—specifies some items that have their intrinsic values in the most elemental way. The *basic intrinsic value states* of each axiology are the items that the axiology takes to be the fundamental bearers of intrinsic value. Each of these items has its intrinsic value in a nonderivative way (Feldman 2004, p. 173; Harman 1967; Michael Zimmerman 2001). Think of each such item as an

"atom" of value. The intrinsic value of a complex thing, such as a life or a world, is a function of—the sum of—the value of these atoms. In addition, Fred Feldman suggests that different answers to the following question distinguish *monists* from *pluralists* in axiology: how many properties are there such that intrinsically good basic intrinsic value states are pure attributions of those properties? Monists answer "one"; pluralists answer "several." One sort of monist, the sensory hedonist, says that the relevant property is the property of *feeling sensory pleasure of certain intensity at a certain time;* one type of pluralist submits that the pertinent two properties are the property of *feeling pleasure* and the property of *knowing* (Feldman 2004, pp. 184–85)

Just as some claim that *feeling pleasure* and *knowing* are atoms of value, so it may be ventured that, in addition to other items, loving, too, is such an atom. We tend to favor a monistic approach of the sort Feldman defends. We believe that a version of attitudinal hedonism is a highly credible version of well-being although we do not substantiate this assumption in this work.[11] Roughly, the attitudinal hedonist claims that your life is going well for you to the extent that it contains more intrinsic attitudinal pleasures than intrinsic displeasures. Such pleasures (and displeasures) should be distinguished from sensory pleasures and pains. A person experiences sensory pleasures at a time if she feels pleasurable sensations then. Attitudinal pleasures are not sensory pleasures; rather, they are propositional attitudes. A person takes attitudinal pleasure in something "if he enjoys it, is pleased about it, is glad that it is happening, is delighted by it" (Feldman 2004, p. 56).[12] And a person takes attitudinal displeasure in something, roughly, if he is averse to it. To take intrinsic pleasure or displeasure in an object is to take pleasure or displeasure in it for its own sake.

To formulate a stock version of the simple theory, we introduce some assumptions. First, the bearers of intrinsic value—the atoms of value—on attitudinal hedonism are states of affairs (Feldman 2004, p. 173).[13] Second, whenever a person takes intrinsic pleasure in something, he takes pleasure of some degree where 'degree' is to be understood as strength of attitude. Third, when a person takes intrinsic pleasure in something, he does so for a period of time. Corresponding things are true of intrinsic displeasures. Thus, there are episodes of intrinsic attitudinal pleasure and displeasure. According to attitudinal hedonism, the atoms of intrinsic value are episodes of intrinsic attitudinal pleasure and displeasure all relevantly like the following:

> *Life Atom of Intrinsic Value:* At noon on Tuesday, October 16, 2006, Bob takes intrinsic attitudinal pleasure of intensity +8 (for 10 minutes) in the fact that Bob's beer is frosty cold.

The theory may now be formulated in this way:

Simple Attitudinal Hedonism

1. Every episode of intrinsic attitudinal pleasure is intrinsically good; every episode of intrinsic displeasure is intrinsically bad.
2. The intrinsic value of an episode of intrinsic attitudinal pleasure is equal to the amount of pleasure contained in that episode, with longer and stronger episodes being intrinsically better. Corresponding things are true about the intrinsic value of an episode of intrinsic displeasure. (Feldman 2004, p. 66)

Simple attitudinal hedonism is compatible with its being the case that the atoms of value that contribute to the intrinsic value of a world ("world atoms") may differ, in significant respects, from the atoms that contribute to the intrinsic value of a person's life ("life atoms"). Let us though, work with a version of attitudinal hedonism which stipulates that world atoms are no different than life atoms. This variety of hedonism supplements clauses (i) and (ii) with a third:

3. The intrinsic value of a life is entirely determined by the intrinsic values of the episodes of intrinsic attitudinal pleasure and displeasure contained in the life, in such a way that one life is intrinsically better than another if and only if the net amount of intrinsic attitudinal pleasure in the one is greater than that sort of pleasure in the other. (Feldman 2004, p. 66)

The simple theory may be modified to accommodate various concerns. For example, suppose Hal's life contains far more intrinsic attitudinal pleasures than intrinsic displeasures but each of the states of affairs in which Hal takes pleasure is false. Some may object that Hal's life should not, contrary to what the simple theory implies, be rated as highly intrinsically good. Hal is, after all, relevantly deceived. He may think that members of his community respect him, and he may take pleasure in the state of affairs, *Hal's being respected by others,* but this pleasure, just like all his others, is a "false" pleasure (Adams 1999, p. 84; Kagan 1998, pp. 34–36; Sumner 1996, p. 98). Although there is room for an attitudinal hedonist to maneuver to contain this worry, perhaps this is one of the simple theory's legitimate shortcomings.

Feldman suggests one response to this sort of objection. *Truth-adjusted intrinsic attitudinal hedonism* adjusts the values of episodes of intrinsic attitudinal pleasure according to whether the pleasures are taken in true objects (but not of intrinsic displeasures for reasons that Feldman discusses; Feldman 2004, pp. 111–14; 181–82). On this adjustment, the fundamental goods are takings of intrinsic pleasure in various states of affairs supplemented with the qualification that such takings of pleasure enhance the value of a life more when they are takings of pleasure in *true* states of affairs. If Hal* lives the very same sort of life that Hal lives—this counterpart takes

intrinsic pleasures (and intrinsic displeasures) to the same degree in the very sorts of object that Hal does—save that Hal*'s pleasures are all pleasures in true objects, the truth-adjusted theory implies that Hal*'s life is intrinsically better for Hal* than is Hal's life for Hal. This truth-adjusted hedonism has it that truth has no independent intrinsic value; rather, truth enhances the intrinsic values of the only things that are the bearers of intrinsic value: atoms of intrinsic attitudinal pleasure and displeasure.

Here is a different concern with the simple theory. Suppose, again, that a person's life contains far more intrinsic attitudinal pleasures than displeasures but this time all his pleasures are pleasures taken in "worthless" objects. G. E. Moore (1903, ch. III, sec. 56) speaks of "a perpetual indulgence in bestiality." We may take Moore and like-minded people to be objecting that this person's life should not, unlike what the simple theory implies, qualify as highly intrinsically good. To meet this objection, Feldman describes a way in which we could adjust the value of a pleasure to reflect the extent to which the object of that pleasure deserves to be enjoyed. The root insight is that intrinsic attitudinal pleasures taken in objects that deserve to have pleasure taken in them are to be rated as intrinsically better than otherwise similar pleasures taken in objects that do not deserve to have pleasures taken in them. Similar adjustments for desert are made in the evaluation of intrinsic displeasures (Feldman 2004, pp. 117–22).

These two objections against the simple theory are objections to what some regard as troubling implications of its ratings of various *lives*. W. D. Ross (1930, p. 138) raises a powerful objection to what he takes to be an unacceptable implication of the theory's evaluation of various *worlds*. It will be instructive to sketch this objection and Feldman's response to it because, as we will see, the response suggests a possible link between love and personal well-being.

Ross imagines two worlds qualitatively identical with respect to what we stipulate is the intrinsic attitudinal pleasure and displeasure contained in the lives of people in these worlds, but different in that in one world, those who receive pleasure are virtuous and those who receive displeasure are vicious whereas in the other, those who receive pleasure are vicious and those who receive displeasure are virtuous. Ross's objection is that the simple theory declares the worlds equally valuable, whereas he judges the former to be better than the latter. We agree with Ross's evaluation of these worlds as does Feldman.

To appreciate Feldman's interesting response to this objection, let's first say a few preliminary things about desert. There are many different desert bases; there are, that is, many different factors that affect the extent to which a given person deserves a certain pleasure or displeasure: excessive or deficient past receipt, moral worthiness or virtue, legitimate claims, established character, etc. Consider virtue (or viciousness). Among other things, a virtuous person is a person who habitually acts "from" virtue; similarly, a vicious person habitually acts "out of" vice. You may be deserving of pleasure if your acts

exemplify virtue—you perform virtuous deeds; you may be deserving of displeasure if your acts express vice—you perform vicious deeds. It seems, then, that, in this fashion, virtuous and vicious deeds (actions or intentional omissions) are "desert bases." One's actions, though, need not express virtue for it to be true that, because of these deeds, one deserves pleasure. You may give alms to the poor in the belief that, in so doing, you do what is morally obligatory. You may well be deserving of moral praise for your deed even if this act does not spring from virtue; equally, you may be deserving of pleasure.

A person can deserve pleasure for many reasons: she may have performed many morally good deeds; she is innocent and maybe innocent people deserve pleasure in virtue of their innocence; she has been deprived of food and she deserves the pleasures of a good meal, and so on. Analogously, a person may deserve displeasure for many different reasons. Needless to say, a person may, at a time, deserve pleasure for some reasons and may also, at that time, deserve displeasure for other reasons or may receive pleasure or displeasure that is undeserved. "Undeserved," it should be cautioned, masks an ambiguity that is worth exposing. It may be taken to mean the same as either "not deserved" or "*deserved* not." Usually, it is taken to mean the latter. We may bring out this distinction by introducing another set of distinctions concerning desert bases. A person has a negative desert base if the person deserves displeasure; a person has a positive desert base if she deserves pleasure; a person has a neutral desert base if she neither deserves pleasure nor deserves displeasure. Suppose a person with a neutral desert base receives some pleasure. It would be correct to say that she does not deserve to receive the pleasure that she receives; there would be no good reason, though, to say that she *deserves* not to receive the pleasure she receives.

Now consider these axiological principles (AXP) that govern the values of episodes of intrinsic attitudinal pleasure (and intrinsic displeasure) that are deserved or not deserved:

> AXP1: Positive desert base enhances the intrinsic value of an episode of attitudinal pleasure (Feldman 1995/1997, p. 163). If one receives the pleasure that one deserves, the value of that episode of attitudinal pleasure is enhanced (the pleasure is made better).[14]

> AXP1 is relevantly analogous to the widely accepted principle that pleasure in the good is intrinsically good. If someone takes pleasure in the good, and this person deserves because of past good deeds, this pleasure, then the intrinsic goodness of such an episode seems to be enhanced by virtue of his getting what he deserves (Michael Zimmerman 2001, p. 220; Chisholm 1986, p. 63; Moore 1903, p. 224; Smart 1973, p. 24; Lemos 1994, p. 74).

> AXP2: Negative desert base mitigates the intrinsic value of an episode of displeasure (Feldman 1995/1997, pp. 164–65). If one receives the

attitudinal displeasure that one deserves, the value of that episode of displeasure is mitigated (the displeasure is made less bad).

One's receiving displeasure is (other things equal) intrinsically bad. Again however, it seems highly plausible that if one receives the displeasure that one deserves, this is not *so* bad; the value of the episode of displeasure is mitigated.

AXP3: Neutral desert base neither enhances nor mitigates the value of pleasure or displeasure (Feldman 1995/1997, pp. 166, 168–69). The intrinsic value of an episode of pleasure (or displeasure) of this sort is directly proportional to the amount of pleasure (or displeasure) it contains.

Suppose you do not deserve any pleasure and you do not deserve any displeasure but you receive some pleasure and displeasure. It appears that the values of the episodes of gratuitous pleasure (and displeasure) you receive mirror the amount of pleasure or displeasure in those episodes.

AXP4: Negative desert mitigates the intrinsic goodness of pleasure. If someone deserves displeasure but gets pleasure instead, the value of that pleasure is mitigated (the pleasure is made less good; Feldman 1995/1997, pp. 164–65).

Finally,

AXP5: Positive desert aggravates the intrinsic badness of pain. If someone deserves pleasure but gets displeasure instead, the value of that displeasure is aggravated (the displeasure is made even worse; Feldman 1995/1997, pp. 166–67).

Let the values of the episodes of intrinsic pleasures and displeasures that a subject experiences be adjusted for desert in accordance with principles AXP1 through AXP5. Feldman dubs this new measure of value *"subject's desert-adjusted intrinsic value."* Subject's Desert-Adjusted Intrinsic Attitudinal Hedonism is the view that the value of *a world* is the sum of the subject's desert-adjusted values of the intrinsic attitudinal pleasures enjoyed and pains suffered in that world. The theory runs as follows:

Subject's Desert-Adjusted Intrinsic Attitudinal Hedonism (*SDAIAH*)

1. Every episode of intrinsic attitudinal pleasure is intrinsically good; every episode of intrinsic displeasure is intrinsically bad.
2. The subject's desert-adjusted intrinsic value of an episode of intrinsic attitudinal pleasure is equal to the amount of pleasure contained in that episode adjusted for subject's desert; the subject's desert-adjusted

intrinsic value of an episode of intrinsic displeasure is equal to the amount of displeasure contained in that episode adjusted for the subject's desert.

3. The intrinsic value of a world is entirely determined by the subject's desert-adjusted intrinsic values of the episodes of intrinsic attitudinal pleasure and displeasure contained in that world, in such a way that one world is intrinsically better than another if and only if the net amount of intrinsic attitudinal pleasure adjusted for subject's desert in the one is greater than the net amount of that sort of pleasure in the other. (Feldman 2004, p. 195)

Consonant with Ross' estimation, Subject's Desert-Adjusted Intrinsic Attitudinal Hedonism rates the just world as intrinsically superior to the unjust world.

9.5.2. A Digression: On the Value of Worlds and Lives

Feldman underscores the point that Subject's Desert-Adjusted Intrinsic Attitudinal Hedonism offers an evaluation of *worlds* unlike, for example, an attitudinal hedonism that adjusts the values of episodes of intrinsic attitudinal pleasures and displeasures for object-worthiness. This latter sort of hedonism, he says, offers an evaluation of *lives* (Feldman 2004, p. 195). According to Feldman, then, world atoms differ from life atoms; "atomism"—the view that there is a uniform set of atoms for the assessment of worlds, lives, the total consequences of actions, and so forth—is false. One may accept the falsity of atomism, yet wonder why it may not also be true that whether a subject deserves (or fails to deserve) a pleasure or displeasure contributes to the value of that person's *life*.

Reconsider the sort of hedonism—"Object-Worthy Hedonism"—that adjusts the value of an episode of attitudinal pleasure (or displeasure) in accordance with whether the object of that pleasure (or displeasure) deserves to be enjoyed. Feldman, as we have recorded, suggests that some may take object worthiness to have an impact on the intrinsic value of lives. Suppose Spike and his twin, Spike*, have no special aesthetic training. Spike takes great intrinsic pleasure in viewing the Mona Lisa; the object of his pleasure is genuinely beautiful and it deserves to be appreciated. Spike* takes otherwise similar pleasure in a copy of the Mona Lisa that he believes is the masterpiece. The counterfeit is not "object worthy"; it is not an object that deserves to be an object of pleasure. Assume that Spike would have displayed the same sort of enjoyment that he displayed in gazing at the real thing had he, instead, been presented with the fake. Assume, further, that corresponding things are true about his twin. Object-Worthy Hedonism implies that the value of Spike's pertinent episode of pleasure is enhanced in virtue of his taking pleasure in an object that is object worthy; and it implies that the value of Spike*'s pertinent episode of attitudinal pleasure is not so enhanced. This is so even if the relevant life segments of these two

individuals are internally indiscernible—the segments containing the "aesthetic pleasures" of each of the twins may "feel" exactly the same "from the inside."

On route to inquiring whether subject's desert affects the value of lives, we first ask whether indiscernibleness of this sort calls into question Object-Worthy Hedonism. The concern is that since object worthiness need have no influence whatsoever on how one's life "feels" from the "inside," object worthiness is not germane to well-being (though it may be germane to the value of worlds). We are inclined to think that this concern is not on target. *Why,* after all, should it be the case that whether some factor, such as object worthiness, makes a difference to how one's life "feels" from the "inside" be the test of whether this factor is pertinent to an evaluation of lives as opposed to an evaluation of worlds?

Perhaps, one might claim, we should distinguish between personal value and impersonal (or ethical) value. Something is personally good for some person if and only if it is good in terms of the welfare or well-being of that person. Ethical value is one type of nonderivative value. Michael Zimmerman explains:

> When Ross and Feldman . . . say that the world in which the virtuous prosper and the vicious suffer is better than the world in which the reverse is true, they are (I believe) looking at matters from an ethical standpoint. It is *ethically* fitting that personal goods and evils be distributed as they are in the first world, *ethically* unfitting that they be distributed as they are in the second world; hence, an ethically sensitive person would, *ceteris paribus,* prefer the first world to the second. (M. J. Zimmerman 2007, p. 429)

Regarding the sort of adjustment to the values of episodes of pleasure and displeasure that Object-Worthy Hedonism recommends, M. J. Zimmerman questions why we should bother with the adjustments in the first place, since Moore's objection from "base pleasures" seems, often, to stem from *ethical* concerns:

> [S]omeone who disapproves of a life full of "low" pleasures is likely, I think, to be questioning the *ethical* value of such a life, rather than the claim that such a life is good in terms of *personal* welfare. But if this is so, then the proper response to the objection is not to absorb it by adjusting one's theory but to reject it outright as misdirected, since it concerns ethical value and not . . . personal value . . . (M. J. Zimmerman 2007, p. 431)

In M. J. Zimmerman's estimation, Moore's objection gives us reason to believe that a life replete with intrinsic attitudinal pleasures taken in unworthy objects may well be low in ethical value but not in personal value, and it is the latter and not the former that is pertinent to well-being.

While we find Zimmerman's reflections highly suggestive, we do not think that they provide conclusive grounds to unseat Object-Worthy Hedonism. Grant the distinction between personal value and ethical value. At issue, really, is whether object worthiness may bear on *personal value* even if it is allowed that object worthiness may bear on ethical value. One need not be confusing the two sorts of value in recommending that the extent to which an object deserves to have pleasure taken in it may have definite impact on personal value.

In sum, Moore advances an objection—the objection from base pleasures—that may reasonably be regarded as an objection against Subject's Desert-Adjusted Intrinsic Attitudinal Hedonism. On one response, Moore's objection misfires; it conflates personal value and ethical value. We prefer to be more charitable: the Moorean objection resurfaces even if one marks the distinction between these two sorts of value. On Feldman's second response, the proponent of Object-Worthy Hedonism construes Moore's objection as addressing personal value. An implication of this second response is that there are "subject-independent factors," roughly, factors that need not in any way influence what the life of a person "feels" to that person from the "inside," that affect well-being. Object worthiness is one such factor. Others have suggested that truth is another such factor.

We may now finally revert to our initial riddle: why should it not be the case that if a subject deserves (or fails to deserve) a pleasure (or a displeasure), this factor contributes to that person's well-being—it contributes to personal value—even if it is acknowledged that it contributes to the intrinsic value of worlds? Feldman submits that the evaluation of worlds and the evaluation of lives make use of different considerations:

> If we want to know how well a person's life is going for him, we want to know the net extent to which he is enjoying things. On the other hand, if we want to know how well things are going in a world, we want to know something about the extent to which people are enjoying good things and suffering bad things, taking account of the extent to which those people deserve to be enjoying the good things and suffering the bad ones. (Feldman 2004, p. 195)

We find the view that different considerations pertain to the evaluation of lives and to the evaluation of worlds attractive. Indeed, there *are* prima facie powerful reasons to believe that desert (or justice) may be welfare irrelevant but world relevant. Suppose Al commits a crime, say a murder, he deserves the displeasure of long-term confinement, and he does in fact suffer such confinement. He experiences displeasure in accord with what he deserves. However, even if we accept that the actual world has more intrinsic value ceteris paribus, owing to this instantiation of justice, than a pertinent counterfactual world in which Al commits the murder, is not caught, and thus does not suffer the relevant displeasures, it is far from pellucid

that Al's life in confinement is better for Al than his life is for him in the counterfactual world. There is, however, still the quandary that we touched upon previously: precisely what makes some factor welfare irrelevant but world relevant? We prefer to err on the side of caution. If we grant that a subject-independent factor, such as object-worthiness, can bear on the evaluation of lives, it is not transparent why we should deny that another subject-independent factor, such as subject's desert, should also bear on the evaluation of lives (even having conceded that it bears on the evaluation of worlds), given principles such as AXP1 to AXP5 and no decisive view concerning the features a factor must satisfy to be welfare irrelevant.

Let's register an assumption:

> *Assumption Value:* If object worthiness can have an impact on the value of episodes of intrinsic attitudinal pleasures and displeasures, and in virtue of this impact can affect well-being, then subject's desert can also have an impact on well-being by way of having an impact on the value of episodes of intrinsic attitudinal pleasures and displeasures.

We believe that this assumption is at least credible.[15]

9.5.3. Love and Intrinsic Attitudinal Pleasure

We may now finally address this question: assuming some version of attitudinal hedonism as the correct axiology for rating lives, how, exactly, does love contribute to well-being? We offer three suggestions.

First, it is undeniable that when we love, and more generally, when we act from love, we typically take intrinsic delight in the fact that we do; we take intrinsic attitudinal pleasure in the fact that we do. Maybe we do so, in the end because of our nature; this is simply the way we are. Perhaps sociobiological considerations account for the fact that, generally, we delight in matters of the heart. Other things equal, the more one's life is imbued with love, the more intrinsically better that life is for one.

It may be objected that the sort of hedonism to which we appeal is straightforwardly false. This is though, highly controversial. In any event, it is hard to accept the view that intrinsic attitudinal pleasure, and thus delighting in love, does not in *any way* contribute to the intrinsic value of our lives. Should some form of pluralism be true, it would again be difficult to believe that the axiology would not recognize appropriate states of attitudinal pleasure (and displeasure) as one species of atom of intrinsic value. Further, we have certainly not ruled out the possibility that love itself may be a primary constituent of certain atoms.

Another objection is that if attitudinal pleasures are atoms of value, we could just as well bring up our children to take delight in living morally, or virtuously, or competitively, or religiously. Why, then, should we single out love? As a preliminary observation, we stress that we have not *singled* out

love. Our view is that love is at least (derivatively) valuable. This is consistent with the stance that other things are (derivatively) valuable.

Second, let's recapitulate the central tenets of Object-Worthy Hedonism. On this version of attitudinal hedonism,

> the intrinsic value of an attitudinal pleasure is determined not simply by the intensity and duration of that pleasure, but by these in combination with the extent to which the object of that pleasure deserves to have pleasure taken in it. More exactly, the value of a pleasure is enhanced when it is pleasure taken in a pleasure-worthy object, such as something good, or beautiful. The value of a pleasure is mitigated when it is pleasure taken in a pleasure-unworthy object, such as something evil, or ugly. The disvalue of a pain is mitigated (the pain is made less bad) when it is pain taken in an object worthy of pain, such as something evil, or ugly. The value of a pain is enhanced (the pain is made yet worse) when it is pain taken in an object unworthy of this attitude, such as something good or beautiful. (Feldman 2004, p. 120, note omitted)

According to Object-Worthy Hedonism, every episode of intrinsic attitudinal pleasure is intrinsically good, every episode of intrinsic displeasure is intrinsically bad, and the intrinsic value of an episode of intrinsic attitudinal pleasure is equal to what we may call the object's "desert-adjusted amount of pleasure" that it contains (similar things are true regarding the intrinsic value of an episode of intrinsic displeasure). Further, on this theory the intrinsic value of a life is entirely determined by the intrinsic values of the episodes of intrinsic attitudinal pleasure and displeasure contained in that life, in such a way that one life is intrinsically better than another if and only if the net object's desert-adjusted amount of intrinsic attitudinal pleasure in the one is greater than the net amount of that sort of pleasure in the other (Feldman 2004, p. 121).

Just as something's *being genuinely beautiful* is a pleasure-worthy object, so it seems, various states of affairs associated with love are pleasure-worthy objects as well. We have in mind such states of affairs as *someone's defending the trust constitutive of love,* or *someone's being concerned with the welfare of the one who is loved,* or *someone's rejoicing in the achievements or successes of the one who is loved,* or *someone's taking pain in the suffering of the one who is loved*.

Third, let us assume something that is admittedly contentious—our *Assumption Value:* desert affects *welfare* (and not merely the value of *worlds*). More fully, other things equal, deserved intrinsic pleasures are better than otherwise similar pleasures that are not deserved; deserved intrinsic displeasures are less bad than otherwise similar pleasures that are not deserved; and such pleasures and displeasures are pertinent to *welfare*. On this assumption, love enhances well-being owing to its association with

desert. To understand this view we re-emphasize, first, that moral worthiness is a desert base; it has an impact on one's desert level. If two persons are alike in all other relevant respects save that the one has been good whereas the other has been bad, the one that has been good has greater desert level; her desert level for pleasure is higher. We have recorded that there are various duties of love. If one fulfills such duties and one is commendable for them, one has been good from love's standpoint; one's degree of worthiness from love's standpoint has been augmented. If moral worthiness can influence one's desert level for pleasure, then it appears that worthiness from the point of view of love should also be something that can impact one's desert level; the more worthy a person is in this respect, the greater that person's desert level for primary intrinsic goods such as intrinsic pleasures.

We note, second, that a number of people have suggested that the amount of time, effort, or work one has invested in acquiring a good can influence a person's desert level relative to that good. If two people, alike in other respects, except that one has worked hard to cultivate a garden, whereas the other has done nothing in the garden, the hard-working gardener has greater investment; she deserves more to enjoy the benefits deriving from the garden (Feldman 1995/1997, p. 162, note 16).[16] Love is frequently manifested in lovable behavior. Further, one can invest a great deal of time and effort in loving others (simply reflect on the parental duties of love!). It would seem, then, that a person who has invested more in love deserves more to enjoy the benefits deriving from the "love-related good." If, acting from love, one has worked very hard to ensure a successful marriage; one deserves more to enjoy the benefits of a happy marriage.

Thus, love may contribute to the intrinsic value of lives in perhaps three distinct ways. We take delight in concerns of the heart; we take delight, for example, in the fact that we care deeply for the well-being of the beloved. Moreover, the objects of episodes of intrinsic attitudinal pleasure concerning matters of the heart are pleasure-worthy objects. Finally (this is more controversial), love may enhance the value of pleasures in virtue of love's association with desert.[17] We have already catalogued various benefits of love (in Section 7.3.1). Add to this *these diverse ways in which love contributes or may contribute to well-being*, the proposal that we have ample reason to educate with the aim of ensuring that our children turn into loving agents, is alluring.

9.5.4. Being Morally Responsible Agents, Being Autonomous Critical Thinkers, and Well-Being

We now briefly address why becoming morally responsible agents and autonomous critical thinkers are vital (but not essential) elements in the good life. With the former, we suggested (in Section 9.4.1) that turning children into morally responsible agents is a prerequisite for admittance into the moral community. In our interactions with others, we assume

that we are dealing with moral agents; certain behavior acquires meaning or becomes intelligible only against the backdrop that we are morally responsible agents. Strawson (1962) goes so far as to suggest that the morally reactive attitudes, elements constitutive in Strawson's view of moral responsibility, are incipient forms of communication. For instance, they are responses to the quality of will—ill will or good will—expressed in a person's conduct, and they are means of conveying to a person who has done wrong that she will have a differential standing in the moral community.[18] Morally responsible agency is, then, a gateway to a way of life that promises rich rewards. To emphasize only one dimension of this life, participating in interpersonal relationships that we value deeply implicates moral responsibility. One mark of friendship, for example, is being willing to forgive. As we noted (Chapter 6, Section 6.3), forgiveness presupposes that the person who is forgiven is blameworthy for some moral wrong. In summary, our general view is the following. We take delight in various activities such as developing and maintaining deep friendships. However these activities implicate moral responsibility (Haji 2003*b*). It is in this way that moral responsibility is a constituent of the good life.

Critical thinking enters the equation in a slightly different fashion. Such thinking opens up one's mind to the myriad "forms of life," the many different possibilities which are such that one may take delight in these possibilities. One may take pleasure in living the life of a virtuous person that Aristotle commends in the early part of the *Nichomachean Ethics,* or in living the life of wisdom—a life committed to philosophical contemplation—that he advocates toward the end of this work, or in living the life of Seneca's Sage, a life eventually marked by *ataraxia*—the condition in which one is not troubled by unruly desires or emotions. The life of Seneca's Sage or the life of a recluse may leave minimal room for moral responsibility, appraisability from love's standpoint, or critical thinking, yet they may well be lives good in themselves for the agents who live them. (To fathom that such a life might be best for one, though, it is important that one be receptive to other possibilities. Being a critical thinker facilitates being so receptive.) This is why we claimed that becoming morally responsible agents, or agents who are commendable and censurable, or agents who are critical thinkers are vital but not essential elements of the good life. They are vital in that it seems that for most of us, the good life *will be* a life imbued with responsibility and love. They are not essential insofar as, for instance, but the life of the recluse may well be next to shorn of these elements, yet a life rich in itself for its agent.

Taking stock, many educators have thought that a fundamental goal of education is to ensure that our children mature into responsible agents, however divergent their views in other respects. But the notion of responsibility educators frequently and uncritically presuppose is moral. We have suggested that it would be fruitful to reflect on the relative importance of varieties of normative responsibility and then modulate our educational

goals accordingly. The importance of these varieties is, partly but pivotally, to be understood in terms of whether, for instance, acting from love or acting from duty contributes, in the manner that we have outlined, to the intrinsic value of our lives. (We do not rule out the plausible view that the importance of these varieties of normative responsibility is to be assessed partly in terms or whether, for example, acting from love contributes to social- or world-betterment.) Similarly, whether other overarching aims of education are to be associated with other factors (such as being autonomous critical thinkers) is vitally to be gauged in terms of the contribution of these factors to the good life for the agent or, to what amounts to the same thing, personal well-being.

At some time during infancy, the child savors an incipient taste of responsibility-level freedom; she takes the first tentative steps toward becoming a moral agent. The path to freedom is ever so fragile, frequently weighed down with the likes of extreme paternalism or responsibility-subversive manipulation. But there is love's shining promise. If guided and raised with love, she may well blossom into a person who herself takes intrinsic delight in affairs of the heart. Thus achieving enjoyment, she will live a life that we can only hope each and every child—and, indeed, each and every person—lives: a life that is highly valuable in itself for her.

Appendices

APPENDIX A

On Other Solutions to the Manipulation Problem

In this appendix, we summarize and evaluate various other approaches to handling the troubling quandary of manipulation (as expounded in Chapter 2). Assessing how it stacks up against prominent rivals constitutes an important, partial defense of our account (in Chapter 3).

A.1. Fischer and Ravizza's Ownership View

In their path-breaking book, *Responsibility and Control: A Theory of Moral Responsibility,* John Martin Fischer and Mark Ravizza (1998) defend a compatibilist theory of moral responsibility. They argue, first, that responsibility does not require metaphysically open alternative possibilities; so causal determinism does not threaten responsibility merely in virtue of eliminating such alternatives. They subsequently develop an account of directional control—what they call "guidance control"—that they propose is the freedom-relevant (in addition to, for instance, the epistemic) condition of responsibility.

Guidance control has two components, neither of which they argue determinism impugns. A distinction is presupposed between the kinds of "mechanism"—roughly, the type of process—that actually causally issues in the agent's behavior and other sorts of mechanism. The reasons-responsiveness component requires that the mechanism that produces the action be appropriately sensitive to reasons. The ownership component demands that the mechanism be the agent's own. Briefly put, an agent has guidance control in performing an action if and only if the action issues from his own, moderately reasons-responsive mechanism (Fischer and Ravizza 1998, p. 86).

Moderate reasons-responsiveness consists in regular reasons-receptivity, and at least weak reasons-reactivity, of the actual-sequence mechanism that leads to the action (p. 89). Reasons-receptivity is "the capacity to recognize the reasons that exist," and reasons-reactivity is "the capacity to translate reasons into choices (and subsequent behavior)" (p. 69). A

defining characteristic of regular reasons-receptivity is that "it involves an understandable pattern of (actual and hypothetical) reasons-receptivity" (p. 71). A mechanism of the agent that issues in the agent's performing some action in the actual world is weakly reasons-reactive if there is some possible world with the same laws in which a mechanism of this very kind is operative in the agent, "there is sufficient reason to do otherwise, the agent recognizes this reason, and the agent does otherwise" for this reason (p. 63).

Fischer and Ravizza acknowledge that it is possible for an agent's actions to issue from a moderately reasons-responsive mechanism whose primary constituents have been induced externally by clandestine manipulation, hypnosis, subliminal advertising, brainwashing, and so forth. Intuitively, in cases of this sort the agent is not morally responsible for the pertinent actions. Such cases impel Fischer and Ravizza to theorize that the way in which the agent's springs of action are acquired has a pronounced bearing on responsibility. Responsibility is thus, they venture, an essentially "historical" phenomenon. Fischer and Ravizza's prognosis is that in these troubling cases, the mechanism that issues in action is not the "agent's own," the agent having failed to take responsibility for it. Reasons sensitivity thus, requires supplementation with the mechanism-ownership component to counteract the challenging peril of acquiring causal springs in a fashion that subverts responsibility. As Fisher remarks, the "reasons-responsiveness itself cannot have been put in place in ways that bypass or supercede the agent—the mechanisms that issue in one's behavior must be *one's own*" (p. 147).

Taking responsibility, measures by which an agent makes a mechanism "his own," involves three elements. First, the agent must regard himself as the source of consequences in the world by realizing that his choices have effects in the world. Second, the agent must see himself as an appropriate candidate for morally reactive attitudes, such as praise and blame, as a result of how he affects the world. Third, the views specified in the first two conditions—that the individual can affect the external world in certain characteristic ways through his choices, and that he can be fairly praised or blamed for so exercising his agency—must be based on his evidence in an appropriate way.

Several constituents of this engaging theory, such as the account of taking responsibility and the issue of mechanism individuation, merit close scrutiny.[1] Here, we focus on Fischer's response to Pereboom's four-case argument. To remind ourselves of the cases, we recapitulate their essential features.

> **Case 1.** Plum is created by neuroscientists who manipulate his deliberations via radio-like technology, "directly producing his every state from moment to moment" (Pereboom 2001, p. 113). Reasoning as a rational egoist, his deadly action satisfies leading compatibilist conditions for responsibility.

Case 2. Plum is created by neuroscientists who program him to weigh reasons for action so that he is often but not exclusively rationally egoistic. In the relevant circumstances, the egoistic reasons are very powerful and he is causally determined to kill for these reasons (Pereboom 2001, p. 113–14).

Case 3. Plum is an ordinary human being, "except that he was determined by the rigorous training practices of his home and community" that "determined his character" (Pereboom 2001, p. 114). The training practices deterministically result in his act of murder.

Case 4. Plum is an ordinary human being. Raised under normal conditions, he is often but not exclusively egoistic. In the relevant circumstances, his egoistic reasons and background conditions deterministically issue in his act of murder (Pereboom 2001, p. 115).

Remember, Pereboom proposes that Plum is not morally responsible for killing White in Case 1 and he thinks that there are no morally relevant differences pertinent to responsibility between any two contiguous cases. It follows from these verdicts that Plum is not morally responsible in the last case describing ordinary upbringing. In addition, Pereboom claims that this sort of argument undermines any compatibilist candidate, and any libertarian candidate not committed to agent causation. This is because Pereboom believes that deterministic causal histories, as well as indeterministic ones not tied to agent causation, are not relevantly different from a manipulated causal history. So if the latter sort of history undermines free action and moral responsibility, so do the former sorts of history (for elaboration, see Chapter 6, Section 6.2). Pereboom claims that only agent causation can accommodate free action but he believes that empirical considerations speak against our being agent causes.

Fischer responds in the following fashion.

> Pereboom basically asks the compatibilist to point to the place (after Case 1) along the slippery slope where responsibility emerges. My answer: there is no such place, as Pereboom suggests. Rather, on a plausible understanding of the case, Professor Plum is morally responsible in Case 1. Thus, there is no impediment to saying that Plum is responsible in Case 4 (and, in general, in the context of causal determinism).
>
> As Pereboom points out, on my view it turns out that Plum has taken responsibility for the manipulation-mechanism; after all, this is the mechanism on which he always acts, and when an individual develops into a morally responsible agent, he takes responsibility for his actual-sequence mechanisms, even if he does not know their details. Further, Pereboom is at pains to point out that the desires on which Plum acts are not irresistible; I take it that Pereboom wants to

> say that there is no psychological (or other) *compulsion* here, but mere causal determination. It follows that Plum acts from his own, moderately reasons-responsive mechanism; holding fixed the actual kind of mechanism, there is a suitable range of possible scenarios in which Plum recognizes reasons to do otherwise and does indeed behave in accordance with those reasons. (Fischer 2004, pp. 156–57).

Fischer explains that his response to Pereboom's challenge involves two factors: (1) the distinction between moral responsibility and other moral ascriptions, such as blame- or praiseworthiness, and (2) the distinction between mere causal determination and action from a compulsive or irresistible desire. Regarding the latter, Fischer says that even if there is direct manipulation of Plum's brain in Case 1, the manipulation does not issue in desires so strong as to count as compulsions. Plum's "actual-sequence mechanism has the general power or capacity to respond differently to the very reasons that actually obtain in the case" (p. 157). Although Plum is manipulated, he is not compelled to act as he does; thus, he is not a robot—he has a certain minimal measure of control with which responsibility is associated (p. 157). As for the first distinction, Fischer writes,

> But it is of course also very important to mark the difference between being morally responsible (in virtue of exercising guidance control) and actually being blameworthy (or praiseworthy). In my view, further conditions need to be added to mere guidance control to get to blameworthiness; these conditions may have to do with the circumstances under which one's values, beliefs, desires, and dispositions were created and are sustained, one's physical and economic status, and so forth. Plum, it seems to me, is not blameworthy, even though he is morally responsible. That he is not blameworthy is a function of the circumstances of the creation of his values, character, desires, and so forth. But there is no reason to suppose that anything like such unusual circumstances obtain *merely* in virtue of the truth of causal determinism. Thus, I see no impediment to saying that Plum can be blameworthy for killing Mrs. White in Case 4. Note that there is no difference with respect to the minimal control conditions for moral responsibility in Cases 1 through 4—the threshold is achieved in all the cases. But there are (or may be, for all that has been said in Pereboom's descriptions) wide disparities in the conditions for blameworthiness. (Fischer 2004, p. 158)

Fischer's response is highly instructive. One of its virtues is that it impels us to clarify the cases, especially the first two. Unless one is already committed to various incompatibilist principles, such as the principle that one cannot be morally responsible for an action if it is produced by a source over which one has no control (Pereboom 2001, p. 4), contingent upon the details of the case, a neutral party to the debate—a party not wedded to

compatibilism or incompatibilism—might arrive at a verdict concerning Plum's responsibility that is very different from Pereboom's verdict. Many of our desires originate in or from external sources over which we have no control. It is not obvious though, that we cannot be responsible for behavior that issues from such desires as long as we have some measure of control in relation to them, such as altering their strength, revising them, or altogether eradicating them. If this is so then it should, similarly, not be *obvious* that we cannot be responsible for behavior that issues from desires which neuroscientists surreptitiously implant in us. A second, related virtue of Fischer's response is that it highlights the fact that though one may not be responsible for action that derives from irresistible desires, one may still be responsible—perhaps to a lesser degree—for action stemming from hard-to-resist but not compulsive pro-attitudes. A third virtue of the response is sensitivity to the view that although one's intuitions may not sit right with the thought that Plum, in Case 1, is responsible for the killing, there is an explanation for why the intuition is more-or-less on target. Fischer's proposal is that one may be conflating moral responsibility with, for example, moral blameworthiness, and overlooking the fact that conditions which moral responsibility require differ from those which moral blameworthiness require. As Fischer says,

> Moral responsibility, as Ravizza and I understand the notion, is more abstract than praiseworthiness or blameworthiness: moral responsibility is, as it were, the "gateway" to moral praiseworthiness, blameworthiness, resentment, indignation, respect, gratitude, and so forth. Someone who is morally responsible is an *apt candidate* for moral judgments and ascriptions of moral properties; similarly, a morally responsible agent is an *apt target* for such attitudes as resentment, indignation, respect, gratitude, and so forth. Someone becomes an apt candidate or target—someone is "in the ballpark" for such ascriptions and attitudes—in virtue of exercising a distinctive kind of control ("guidance control"). But it does not follow from someone's being an apt target or candidate for moral ascriptions and attitudes that any such ascription or attitude is justifiable in any given context. After all, an agent may be morally responsible for morally neutral behavior. Further, an agent can be morally responsible, but circumstances may be such as to render praise or blame unjustifiable. (Fischer 2004, pp. 157–58; note omitted)

However, there are various concerns with Fischer's response. First, in the passage just cited, Fischer claims that someone who is morally responsible is a suitable candidate for moral judgments and ascriptions of moral properties. Some caution, though, is called for. Imagine that Jones is the lead character in a so called "Frankfurt-type example" after Harry Frankfurt's (1969/1988) development of the example. The example includes an

arrangement that supposedly ensures that Jones has no pertinent causal alternatives without in any way influencing Jones' behavior, thus permitting the behavior to be "truly Jones' own" without being "up to Jones." In Frankfurt's original version, Jones seemingly has various alternatives from which to choose. He makes the choice and performs the action in question not knowing that, had he revealed even the slightest inclination to act differently, something, a "counterfactual intervener," would have forced him to act as he in fact did. Since he acted on his own in that the intervener did not play any role in his action—it is just that the intervener would have forced him to act as he did had he shown any signs of acting differently—it seems that he should be morally responsible for his deed. He should be so even though he "could not have done otherwise." In this Frankfurt-type situation, though Jones is blameworthy and, hence, morally responsible for his pertinent action, Jones' action will *not* be morally obligatory, right, or wrong, because obligation, right, and wrong require alternative possibilities.[2] So despite Jones' being morally responsible for his deed in his Frankfurt-type situation, it is false that, in his situation, he is an apt candidate for moral judgments of obligation, right, or wrong.

Second, we should bear in mind the distinction between overt blame (or praise) and blameworthiness (or praiseworthiness; Haji 1998, pp. 9–10). The former is concerned with the outward expression of blame by words, gestures, or actions, the latter with being deserving or worthy of blame. It is puzzling how one's economic status, as Fischer suggests in the passage cited above, is associated with a condition of blameworthiness but less enigmatic how it may be affiliated with overt blame. For instance, it may be unfortunate but nonetheless morally required that a person of significant economic stature in society not be overtly blamed for something for which she is blameworthy. Pereboom's concerns are with blameworthiness and not with overt blame.

Third, like moral responsibility, blameworthiness has several requirements including epistemic and control requirements. Regarding the latter, in the absence of good reason to believe otherwise, blameworthiness and responsibility, it appears, have the very same requirements. We may, thus, suppose that Plum's action of murdering White satisfies these freedom or control requirements in all four cases.

With respect to the epistemic requirement, some claim that a condition of moral blameworthiness is that the agent know that her action in question is morally wrong. But this would be too stringent. A Frankfurt-type case, among other things, tells against this condition. It is much more plausible that blameworthiness requires belief in what is wrong. Specifically, an agent is blameworthy for performing an action only if the agent performs it (at least partly) on the basis of the belief that he is doing wrong in performing it (Haji 1998, ch. 9). Some insist, in addition, that moral blameworthiness requires moral wrongdoing: one is blameworthy for an action only if it is wrong for one to do it. We have serious reservations about this condition

(Haji 1998, ch. 8). However even if it is accepted, we can plausibly suppose that Plum's action of murder in each of the four cases satisfies this condition and that Plum in each of these cases meets the previous doxastic condition.

Fischer suggests that blameworthiness "may have to do with the circumstances under which one's values, beliefs, desires, and dispositions were created and are sustained" (Fischer 2004, 158). We agree but this is also true of moral responsibility. If manipulators were to implant irresistible desires in Plum to kill White, or Plum were to acquire such desires as a result of strange forces impinging on him while traversing the Bermuda Triangle, the conditions under which the desires were "created and sustained" would, presumably, make a difference to assessments of responsibility. Now, of course, we do not deny that the manipulators can arrange things to ensure that though Plum exercises the control that moral responsibility and blameworthiness require, Plum is responsible—he is an apt target for the reactive attitudes in relation to what he does—but *not* blameworthy for what he does. The manipulators may, for instance, instill in Plum the belief that it is not wrong for him to kill White, indeed, they may instill in him the belief that it is obligatory, and they may see to it that Plum sustains this belief over time. Nothing about the four cases *requires* that they be construed in this way.

Summing up, our position is the following. Blameworthiness, just like moral responsibility, has control, epistemic, and authenticity requirements. Sometimes, a person may be morally responsible for performing an action—in the sense that he is an apposite candidate for the reactive attitudes on the basis of performing that action—but he may not be blameworthy for the action. Suppose we interpret Case 1 in a fashion in which Plum *is* morally responsible for killing White. In this event, we do not see why Case 1 is *not* also amenable to the construal that Plum is blameworthy. In short, we do not see the asymmetry in ascriptions of responsibility and blameworthiness in the four cases that Fischer discerns. This, in turn, imperils what we have proposed is Fischer's suggested explanation of why the intuition of many that Plum is not morally responsible in the first case is roughly on track.

Finally, suppose it is conceded that Plum is *not* morally blameworthy for killing White in Case 1. For all Fischer has said, it is not clear why Pereboom's challenge does not resurface at the level of blameworthiness: Pereboom might ask the compatibilist to point to the place (after Case 1) along the slippery slope where blameworthiness emerges. With due fairness to Fischer, perhaps much more requires to be said about the further conditions needed "to be added to mere guidance control to get blameworthiness" (Fischer 2004, p. 158) to see whether this challenge can be met by distinguishing between responsibility, on the one hand, and blameworthiness, on the other.

A.2. Yaffe's Tracking Approach

In a recent piece, "Indoctrination, Coercion and Freedom of Will," Gideon Yaffe (2003) seeks to explain why manipulation (of relevant sorts) sometimes

undermines the freedom responsibility requires when other causal forces that produce just the same results as the manipulation do not adversely affect freedom. Yaffe distinguishes two kinds of manipulation. He says that in both, the victim sees herself as having most reason to do what the manipulator wants her to do regardless of whether she is aware of what the manipulator's designs on her are (Yaffe 2003, p. 340). With the first kind of manipulation, indoctrination, the manipulator causes the victim to respond to reasons in a way advantageous to the manipulator. With the second kind, coercion, the manipulator causes the victim to have predominant reason to do only what the manipulator wants the victim to do. To develop these views, Yaffe disambiguates "reasons." The first sense of "reasons" pertains to explaining action. Beliefs that constitute reasons in this sense, "rationales" as Yaffe says, can be false. You may press the red button on the vending machine in light of the false belief that in so doing, a can of Guinness will be dispensed. In the second sense of "reasons," a person has reasons for performing an action "just in case the performance of that action would actually be good in some way or another" but not necessarily overall good (pp. 341–42). Reasons in this sense "confer some degree of legitimacy on the actions they favor" (p. 341), and it is this conception that is of concern in exploring the freedom-undermining effects of manipulation.

Regarding the first kind of manipulation, victims of indoctrination evince a new pattern of taking facts to be reasons for acting in particular ways. The pattern has both a reactive and receptive component. The former concerns the recognition of reasons—what one takes to be reasons. The latter concerns the response to reasons, including translating what one takes to be reasons for conduct into choices and behavior. To facilitate exposition of his views, Yaffe invites us to call an agent who is disposed to evince a particular pattern of responsiveness to reasons as a result of manipulation, "the Manipulated." He has us call another agent, who is just the same as the Manipulated, save that a causal force, *indifferent* to what pattern of reasons she evinces but happens, nevertheless, to produce in her a choice-engendering mechanism ("the set of dispositions underlying the relevant beliefs and desiderative attitudes") with just the features as that fashioned in the Manipulated, "the Unlucky" (p. 343). Yaffe writes,

> The crucial fact about a manipulator who aims to produce in you a certain pattern of response to reasons is that he tracks the production in you of that pattern of response. It is true of the Manipulated, and not of the Unlucky, that were the Manipulated to stray in some way or another from coming to have dispositions to recognize and respond to reasons of the sort that the manipulator wants her to have, he would take steps to see to it that she was placed back on course. The Unlucky, on the other hand, would simply stray away from the course and come to have a different pattern of response to reasons from that of the Manipulated. (pp. 343–44)

Yaffe explains that when we fall into the hands of indoctrinators, fewer lives—fewer "kinds of pattern of response to reasons"—are available to us, given our unchangeable features and "fixed" crucial aspects of our past; fewer lives are available to us than are available when we are the unlucky victims of neutral causal forces (p. 345). It is in this crucial way that indoctrination constrains freedom.

Yaffe emphasizes that, unlike some other accounts of the freedom-limiting impact of indoctrination, the danger to freedom from indoctrination derives from being in the hands of something that *tracks* the production in the agent of a particular kind of tendency or mechanism to respond to reasons. While "such tracking is often associated with the presence in the tracker of certain desires or intentions, what matters for freedom is the tracking, not the desires or intentions" (p. 347). A robotic tutor, for example, just like a real tutor, may indoctrinate the "victim" partly as a result of tracking results (pp. 346–47). If someone or thing, such as a neutral causal force, produces in an agent tendencies for response to reasons without tracking the production in the agent of these tendencies, the agent suffers damage but not the kind distinctive of manipulation. Unlike indoctrination, such interference does not limit the range of possible lives the agent might have led.

Since it is Yaffe's views on indoctrination that are of primary concern, we can afford to be relatively brief with his account of coercion. Yaffe proposes that whereas freedom-undermining indoctrination produces in its victim a particular mechanism for response to reasons, coercers capitalize on the mechanism for responding to reasons that their victims already possess. Coercers endow their victims with reasons for acting as the coercers want them to act. Yaffe says,

> The key to the explanation for the freedom-undermining force of coercion is that, as a general rule, coercers don't merely produce, but also track, the compliance of their victims. A robber who threatens to injure the cashier should the cashier not hand over the money would usually be ready to threaten a more serious injury were the cashier to prove unresponsive to the first threat. That is, the coercer is rarely attached to the particular nasty consequence that he threatens; with some limits, he is ready to bring about whatever consequence would serve to bring the victim around to compliance. . . . The coercer tailors his threat to the features of the mechanism for response to reason that the victim possesses. (p. 351)

Cases of coercion, Yaffe claims, that do not involve tracking the compliance of the victim in the way described do not diminish the victim's freedom.

In sum, according to Yaffe, indoctrination causes another person to respond to reasons in a pattern that serves the manipulator's ends; coercion supplies the victim with reasons that, given the pattern in which he

responds to reasons, move him to act in ways that serve the manipulator's ends. Both forms of manipulation undermine freedom because, unlike neutral causal forces, manipulators track the compliance of their victims.

As a preliminary comment on these suggestive views, Yaffe submits that cases of coercion that do not involve tracking the compliance of the victim in the way described—issuing ever more effective threats to ensure that the victim serves the manipulator's ends—do not diminish the victim's freedom. Certain cases though, in which the effects of coercion are simulated in the absence of a purposeful coercer suggest that a kind of freedom *is* attenuated even when there is no relevant tracking. Michael Zimmerman introduces this case: "a customer enters a bar, hears a menacing voice behind him say, 'Don't turn around! Raise your hands! One false move and you're history!,' and fearfully obeys—only to discover that the voice came from a television in the corner" (M. J. Zimmerman 1988, p. 105). Zimmerman suggests that the customer was strictly free not to give in to the compulsion but that he was not broadly free. Similar things are true of the coerced bank teller who, responding to the threat, "the money or your life!," strictly freely handed over the money but was not broadly free to fail to comply with the coercer's demands.

Yaffe's account appears to generate intuitively unacceptable results in cases involving global manipulation or, perhaps, fails to speak to these cases at all. Imagine a scenario, somewhat analogous to Psychohacker, in which Max is indifferent to the pattern of responding to reasons that is produced in his victim as long as he believes that some deviant pattern *is* produced and in which there is no tracking. Max's victim, Rupa, ends up with the psychological profile of Beth, the benign philosophy professor. Her engineered-in pro-attitudes are practically unsheddable. Intuitively, Rupa is a victim of something approaching indoctrination or coercion—she is manipulated—but Yaffe's views do not seem to have this implication. Unbothered about whether, or believing that, the global transformation is successful in this particular instance, Max forbears checking to see whether Rupa has been suitably transformed and, hence, forbears initiating corrective steps if Rupa is not so transformed. On Yaffe's views, without such tracking, there is no coercion or indoctrination.

One may object that the global manipulation *does* limit the range of possible lives Rupa might have led. But it is controversial whether the range is limited in the manner in which Yaffe's conception of manipulation requires that it be limited. We are, after all, concerned with the effects of manipulation. Global manipulation alters a person in a way analogous to the way in which a radical religious conversion might alter a person. In either case, the agent undergoes a sort of moral death and is "born again." Suppose Ralph undergoes radical religious conversion, not because of indoctrination, coercion, or the like but as a result of the activity of "neutral causal forces." It appears to be false that fewer lives are available to globally transformed Rupa than are available to transformed Ralph who is a "victim" of the neutral

forces. We can imagine that, once transformed, Rupa has all the lives available to her that are available to her psychological twin, Beth.

Still, one might worry that transformed Rupa cannot live the life that she formerly lived, and the range of lives available to her now is different than the range of lives that would have been available to her had she not fallen victim to Max. It is, though, also true that, given the neutral forces to which Ralph is now subject, he cannot live the life that he formerly lived, and that had different neutral forces influenced Ralph, the effects of which would have resulted in his not having undergone the religious change, he would have had available a different range of possible lives than he now has. However, Ralph is not, solely for this reason, now relevantly constrained in the availability of possible lives. Then, too, transformed Rupa should not be relevantly constrained in the availability to her of possible lives.

A second troubling implication of Yaffe's views is that education, even with an eye toward turning children into critical thinkers, is indoctrinative. We accept Siegel's plausible recommendation that a critical thinker is disposed to acquire and maintain beliefs on the basis of evidence for them (Siegel 1988). Accordingly, with the aim of ensuring that our children develop into critical thinkers, the educator seeks to produce in the child a certain pattern or mechanism of response to reasons: a goal of the educator is that the child acquire and hold *evidentially based* items, such as beliefs or values, that qualify as reasons for action; the child should be both receptive and reactive to such reasons. In addition, the educator tracks the production in the child of a pattern of reactivity to such reasons. If the child strays "in some way or another from coming to have dispositions to recognize and respond to reasons of the sort that the . . . [educator] wants her [the child] to have [reasons involving, for example, beliefs that are evidentially held], he would take steps to see to it that she was placed back on course" (Yaffe 2003, p. 343). Such training, if successful, limits the range of possible lives available to the child roughly, to those lives in which it is false that central beliefs are non-evidentially acquired and held. So, on Yaffe's conception of manipulation, education in the service of critical thinking is indoctrinative.[3]

Yaffe entertains the view some endorse that agents who have been induced to respond to reasons in a way superior to the way in which they would have responded in the absence of manipulation are *liberated;* their freedom is not curtailed. If this position can be sustained, then, arguably, children educated to become critical thinkers, are not victims; they have been induced to respond to reasons in ways superior to the ways in which they would have responded in the absence of their training. However, the position is not without shortcomings. Yaffe says that to "accept this line of thought is to see freedom of will as constituted by correct responsiveness to value. It is to insist that a causal influence on one's choices takes away from one's freedom only if there is something bad about having one's choices so influenced"(p. 348). Intuitively though, it appears that freedom can be compromised even if there is something good and nothing bad about

manipulation influencing one's choices. The globally manipulated agent who, as a result of the manipulation, is transformed into a saintly person is still a victim of manipulation. (We overlook the complication that the precise content of the claim, there is *something bad* about one's choices being causally influenced by, say, compulsion, is not pellucid. One may insist that there is "something bad" about *any* form of manipulation, even inept manipulation.)

The view that correct responsiveness to value constitutes freedom of the will, some may claim, might garner support from theoretical considerations. The best account of the freedom responsibility requires, they may venture, implies this view. So, for instance, if the freedom responsibility presupposes is the freedom to act in accordance with the right and the good, as Susan Wolf (1990) proposes, then free will would be constituted by correct responsiveness to value. However, the proof is in the delivery. Wolf's view, for instance, is not without difficulty.[4]

Yaffe (2003) says that his explanation for the freedom-undermining effect of indoctrination

> does not require the claim that the mere fact that fewer ways of responding to reasons are available to an agent is sufficient for her freedom to be diminished. Rather, what is claimed is that when we ask what kinds of lives were available to such agents we find that there were fewer available to the indoctrinated than to those who have come to be as they are through neutral causal forces. If we think that this question does not apply to the agent before us [assume that the agent is the child turned into a critical thinker], for some reason or another, then we are not likely to see her manipulation as freedom-undermining even if fewer ways of responding to reasons are available to her as a result of manipulation. But this is precisely what someone who sees improvement though manipulation as having no damaging effect on an agent's freedom holds. For someone who thinks of freedom this way, of agents improved by manipulation the question of unfreedom never arises, for they have been given a gift by the manipulator rather than twisted into something imperfect. (pp. 348–49)

An assumption of the proponent of the view that manipulation (or, to use more neutral terminology, "interference") that improves the agent has no damaging effect on that agent's freedom seems to be this: the interference is not indoctrinative if it rules out an undesirable pattern of responding to reasons. It would follow that the interferences required to turn children into critical thinkers are not indoctrinative because they preclude undesirable patterns of responding to reasons; if properly educated, the maturing child will not respond to reasons that, for example, involve beliefs non-evidentially acquired. However the assumption is questionable. Gully acquires his values owing to subtle indoctrination on the part of his pastor who has

his own agenda: the pastor wishes to ensure that the conduct of his "flock" will bring sizable returns to him. By sheer coincidence, though, the moral values Gully and others acquire as a result of the pastor's wielding his skills, turn out to be the true or correct values (we suppose that "value nihilism" is not in the running). Assume that the cult leader's machinations are so effective that members of his flock cannot but respond to the "true and the good." It still seems, though, that the pastor's teachings have a damaging effect on Gully's freedom; Gully *is* a victim of indoctrination.

Faced with these problems, we conclude that an alternative to Yaffe's account may fare better.

A.3. David Zimmerman on Substantive Preference-Acquisition

David Zimmerman has lately addressed what he calls the "puzzle of naturalized self-creation in real time": how does the heteronomous child eventually develop into an autonomous person (D. Zimmerman 2003*a,* p. 638)? The conception of autonomy of interest is a conception that essentially involves moral responsibility or accountability. D. Zimmerman claims that the puzzle is particularly pressing for compatibilists who are *positive source historicists* because such compatibilists must explain how, if at all, this transition occurs in a causally determined world. The source historicist holds that the conditions of autonomous or responsible agency include facts about how the person acquired her responsibility-grounding psychological properties, such as reflecting upon what to do or forming higher-order endorsing or rejecting volitions, during her distal history in the external world (p. 642). Opposed to such historicists are *internalist structuralists* who claim that the conditions of autonomous agency are limited to features internal to a person's psychology "during the period of deliberation that proximally precedes action" (p. 642). Negative source historicists either provide a list of motivational springs that are (intuitively) inauthentic, springs affiliated with such things as serious psychoses, deep neuroses, substance addiction, pertinent forms of conditioning, post-hypnotic suggestion, and neurological tampering, or advance an "exclusionary principle" of conditions under which, for instance, the acquisition and continued possession of a pro-attitude is inauthentic (p. 646). For example, Mele recommends that the acquisition and continued possession of a pro-attitude *P* counts as compelled (and so non-autonomous) only if an agent *S* comes to possess *P* in a way that bypasses *S*'s (perhaps relatively modest) capacities for control over his mental life (Mele 1995, pp. 166, 172; we discuss this view in the next section). Positive source historicists give an account of how some but not other heteronomous children manage to become autonomous adults. The negative source historicist insists that desires acquired as a result of intuitively autonomy-subversive manipulation are not "truly the child's own"; the type of manipulation at issue is included on the list of disqualifiers or, presumably, it is of the sort that satisfies the conditions of non-autonomy

stated in whatever exclusionary principle is advanced. The positive source historicist differentiates "stories of 'naturalized self-creation' from sadder ones [for example, those involving autonomy-subversive manipulation] in which autonomy never develops from childhood heteronomy" (p. 647). Such historicists aim to give a principled difference "between the kinds of early preference-acquisition that eventually lead to autonomous agency and those that block the child from transcending its early and inevitable heteronomy" (p. 655). We can regard positive source historicists as, thus, wishing "to clarify the difference between patterns of psychological development that a good liberal would praise as 'education' or 'cultivation,' on the one hand, and condemn as 'indoctrination' or 'psychological manipulation,' on the other" (p. 647).

D. Zimmerman proposes that positive source historicists assume the burden of giving an account of the content and genesis of a number of distinct psychological elements that "ground the development of a child's eventual capacity for autonomous agency" (p. 652). These include concepts, beliefs, procedural commitments, and substantive preferences or evaluations (p. 652). D. Zimmerman's discussion on when substantive preferences are authentic (and when not) is especially revealing. To deal with this central aspect of the puzzle of naturalized self-creation, D. Zimmerman appeals to Richard Brandt's informed preference theory of the good (Brandt 1979). The intuitive core idea, Brandt, among others, develops is straightforward: some desires are inauthentic or "mistaken" because grounded in errors or ignorance. Brandt explicitly addresses the *rationality* of preferences or desires. His official account of the rationality of these pro-attitudes is nonhistorical "in that a preference counts as rational for him if and only if it would survive contemporary 'cognitive psychotherapy,'" (D. Zimmerman 2003*a,* p. 658), such therapy consisting in appropriate exposure now to the "facts and logic" (Brandt 1979, p. 10). However, D. Zimmerman indicates that Brandt's "underlying theory of rational preference is *historicist* through and through, in being 'based on the theory of the *genesis* of pleasures and desires' [Brandt 1979, p. 110; emphasis added by Zimmerman]" (D. Zimmerman 2003*a,* p. 658); and that Brandt, thus, devotes a key part of his project to an account of the ways in which preferences can be mistaken or inauthentic. D. Zimmerman explains,

> The historicist aspect of Brandt's theory comes out in his explanation for *why* certain preferences are extinguishable by contemporary cognitive psychotherapy and thus count as irrational or inauthentic. He remarks that "The production of . . . intrinsic desires and aversions is artificial [inauthentic] *if they could not have been brought about by experience with actual situations which the desires are for or the aversions against*" [Brandt 1979, p. 117; emphasis added by Zimmerman]. Brandt cites four types of "mistakes" that can generate inauthentic preferences:

1. having false beliefs about instrumental consequences.
2. misgeneralizing from untypical examples via classical conditioning.
3. acquiring preferences by means of cultural reinforcers involving social status, and
4. acquiring desires with an overweening strength that is traceable to early deprivation of the object of those desires.

> In each instance the explanation of artificiality or inauthenticity is more-or-less the same: The person would not have acquired the intrinsic preference *but for* having undergone a process of attitude-acquisition that involved some kind of epistemic mistake about the object of the preference. Moreover, the mistake in each instance is more-or-less the same: during the conditioning process the person confuses the *actual* content of the acquired preference with the content of an *extraneous* reinforcer. (pp. 658–59)

D. Zimmerman claims that this sort of historicist account of errors in preference-acquisition developed along Brandtian lines provides the positive source historicist with a promising way to think about the difference between education and indoctrination "because it gives the constraint of evidence-sensitivity some purchase upon the acquisition of substantive preferences" (p. 660). We may regard the proponent of this account as advancing the following criterion of preference inauthenticity.

> *The Historicist Criterion:* If agent *S*'s acquisition of a pro-attitude, such as a desire, involves a relevant epistemic mistake of *S*'s (mistakes of Type 1 to 4, for instance), then that pro-attitude is not authentic.

D. Zimmerman reports that Brandt shies away from this sort of historicist criterion because it involves classical conditioning (the second type of epistemic mistake) that, in turn, implies counterintuitively, that "all acquired preferences, including those as basic to the grounding of autonomous morality as *benevolence* itself, must count as irrational or inauthentic" (p. 661). To elucidate, Brandt distinguishes between empathy, the belief-unmediated response to another infant's expression of pain, and benevolence or sympathy, the belief-mediated aversive response to the same kind of stimulus. Brandt worries that the historicist criterion implies that sympathy is epistemically compromised and so inauthentic, because "the best explanation of its acquisition builds upon the baby's earlier acquisition of empathy via classical conditioning, a process fraught with epistemic error [Brandt 1979, p. 141], and because the best explanation of the baby's acquisition of sympathy upon that foundation also involves classical conditioning [Brandt 1979, p. 144]" (p. 662).

D. Zimmerman argues that Brandt's reservations about the supposed counterintuitive implications of the historicist criterion are exaggerated.

Again, confining his attention to the acquisition of preferences for benevolence or sympathy, Zimmerman proposes that even if the acquisition of these preferences involves classical conditioning, their acquisition may well be free of the relevant epistemic mistakes. The details of Zimmerman's intriguing discussion need not detain us. One major theme of Zimmerman's is that, given the state of cognitive and psychological development of the child, the acquisition of various preferences cannot involve certain varieties of epistemic errors that concern Brandt. A second theme concerns innate preferences. Martin Hoffman's account of infant empathy (Hoffman 1976), upon which Brandt heavily relies, stresses the innate aspects of empathy. Interestingly, Brandt counts innate preferences as rational. Hoffman's views suggest that "the disposition to react empathetically is innate, while only the manner in which it comes to manifest itself is acquired via conditioning" (D. Zimmerman 2003*a,* p. 662). Elsewhere, Zimmerman notes that one cannot simply assume that the autonomous person's responsibility-grounding psychological properties must be acquired for they could be innate (D. Zimmerman 2002, p. 213, n. 41).

D. Zimmerman concludes that there is a strong prima facie case for the rationality and authenticity of preferences for benevolence. He cautions that his discussion centers on whether the processes involved in the acquisition of benevolence contain epistemic mistakes. He stresses that "A compatibilist source-historicist who puts this picture to use in the dialectic must concede that very young children have little control over the formation of their early attitudes, whether the processes that ground them involve nature or nurture, and if the latter, whether they involve epistemically legitimate or illegitimate stimulus generalization" (D. Zimmerman 2003*a,* p. 664).

By way of assessment, the issue of whether the child has control in acquiring and possessing the pro-attitudes constitutive of her initial "evaluative scheme" cannot be underscored.[5] Assume that Youngster's acquisition of central pro-attitudes involves no relevant epistemic errors. If Youngster has little or no control in their acquisition, presumably incompatibilists will not grant their authenticity. One deep incompatibilist concern, after all, can be put in this way. Pro-attitudes acquired as a result of various forms of manipulation, such as indoctrination, are not authentic because they ultimately derive from sources over which one has no control. Would it not, thus, also be true that, if Youngster has no control in the acquisition of the pertinent pro-attitudes, then even if their acquisition is free of epistemic errors, the pro-attitudes are not authentic as they, too, ultimately derive from a source over which Youngster has no control?

We believe that a serious problem afflicting the strategy to secure the authenticity of pro-attitudes D. Zimmerman suggests is the following. As we have formulated it, the historicist criterion is an instance of an exclusionary principle. Like Mele's "bypass principle," it tells us when a pro-attitude is not authentic. *Positive* source historicists, though, want to do more than their negative counterparts. They wish to advance principles

concerning when pertinent pro-attitudes are authentic. Let us then, assume that their position is the stronger one that if the child's acquisition of a pro-attitude, at a stage of the child's development where she has only rudimentary, even if that, cognitive and psychological capacities, does not involve relevant epistemic errors, then the pro-attitude in possession of the child is autonomous or authentic. But this position is untenable. Suppose Youngster acquires a desire to subject central species of belief to rational scrutiny as a result of a learning technique that involves conditioning but no pertinent epistemic errors. It is possible that this mode of acquiring the desire assures that Youngster *cannot but* subject the relevant types of belief to rational scrutiny; the desire is irresistible. He, thus, lacks the control responsibility requires in subjecting the beliefs to such scrutiny. It is plausible to regard this desire as inauthentic. Or reverting to one of Zimmerman's broader themes, suppose Youngster's acquisition of various desires do not involve any epistemic errors because, in part, committing such errors presupposes a degree of cognitive sophistication Youngster lacks. Still, nothing about the acquisition of these pro-attitudes ensures that later behavior to which they give rise will be behavior that respects epistemic or control requirements of responsibility. We suggest that the pro-attitudes are thus, not "truly Youngster's own." Or finally, assume that Youngster is in possession of an innate desire that is irresistible. Then again, Youngster will not be responsible for later conduct that issues from this innate desire, and so the desire it seems, is not authentic. It may be true that generally, victims of compulsive, phobic, or similarly "defective" preferences have acquired such tainted preferences during the course of a history of attitude-formation fraught with the kinds of error that worry Brandt in his rejection of an historicized account of authentic preference (D. Zimmerman 2003*b,* pp. 379–80). However, cases of the sort in which a child acquires, for instance, an irresistible desire via processes that do not run afoul of the germane sorts of error are surely possible. This is all that is required to motivate the concern at issue.

A.4. Alfred Mele's Externalist Historicist Principle

As D. Zimmerman emphasizes, incompatibilists press compatibilists hard on the problem of CNC control or manipulation. Some incompatibilists who are libertarians claim that compatibilist accounts of responsibility, freedom, or autonomy lack the resources for distinguishing at causally deterministic worlds cases in which another agent's control of one's mind victimizes one from cases in which one acts autonomously. The libertarians' challenge to compatibilists may be construed as the challenge to produce an account or set of conditions, compatible with determinism, that allows a principled distinction between actions that are the product of a sequence of causes resulting from CNC manipulation and actions that are the product of a deterministic chain of causes free of such manipulation.

Partly in response to this challenge, Mele (1995, p. 187) advances the following intriguing condition that he proposes is sufficient for psychological autonomy.

An agent, *S*, is (psychologically) autonomous if:

0. *S* is an ideally self-controlled (and mentally healthy) agent;
1. *S* has no compelled* motivational states (where compulsion* is, roughly, compulsion not arranged by S), nor any coercively produced motivational states;
2. S's beliefs are conducive to informed deliberation about all matters that concern him; and
3. *S* is a reliable deliberator.

The condition requires some explanation. Self-control, Mele argues, is not enough for autonomy. To the basic condition of being a mentally healthy agent who frequently and effectively exercises self-control in all domains of his life (condition 0), Mele adds what he calls, the *compatibilist trio* (Conditions 1, 2, and 3). Each member of the trio expresses, respectively, a requirement concerning motivational states, informational (or doxastic) states, and executive processes. Mele proposes that these conditions *do* explain the difference between nonautonomous mind-controlled agents and autonomous causally determined agents, because satisfaction of the compatibilist trio excludes from the psychologically autonomous agents who are victims of CNC manipulation or of nonautonomous mind-control.

Internalist compatibilists hold that a person is autonomous solely in virtue of facts internal to her psychology, such as facts concerning Frankfurtian decisive wholehearted identification (Frankfurt 1987/1988). Externalist compatibilists, in contrast, insist that psychological autonomy hinges on more than what goes in a person's head. Such autonomy requires that the values, desires, beliefs, and so forth that guide self-reflection, deliberation, and action causally depend in the right way on factors in the external world; they must have the right sort of causal history (Fischer and Ravizza 1998, ch. 7).[6]

Responding to the libertarian's challenge, Mele rejects internalist compatibilism (1995, pp. 149–56, 172–73). On a prominent incarnation of this latter view, the effectively exercised ability for critical reflection and self-identification is sufficient for autonomy (Frankfurt 1971/1988; 1987/1988). The autonomy of psychological elements is a non-historical, internal matter in that causal origin and genesis of these elements are irrelevant to their autonomy. A person's reflective identification with, for example, his first-order desires, suffices for the autonomy of these desires. Against such internalism, Mele argues that there is no significant *internal* difference between victims of mind-control or neurological tampering and autonomous agents. "Psychological twins"—globally manipulated Beth and idealized Manson,

for instance—"may be such that only one of them is morally responsible for, and autonomous regarding, the current constitution of his or her 'evil' character" (1995, p. 172). Hence, Mele proposes, the compatibilist trio—especially externalist historical Condition 1—is required to mark the distinction between CNC manipulated and autonomous agents.

On an incompatibilist reading, the problem of CNC manipulation crystallizes to the worry that if CNC manipulation is autonomy-thwarting, causal determination *as such* is autonomy-thwarting, too. On a compatibilist reading, this problem amounts to the worry that internalist autonomy does *not* exclude autonomy-undermining CNC mind-control. Mele tries to dissipate both worries by an appeal to externalist historicism—an appeal to the causal origin and genesis of the psychological states and processes involved in autonomy. The etiology of autonomous states and processes is claimed to be significantly different from that of manipulated and thus nonautonomous states and processes. Melean externalist compatibilist autonomy promises not only to exclude CNC mind control but also effectively to capture the distinction between autonomous agents and CNC manipulated victims despite the fact that both types of agent are causally determined.

Only Condition 1 of the four conditions for autonomy enumerated is explicitly externalist. Even victims of CNC manipulation may satisfy the other conditions. Condition 0 can be met (by *S*) even though *S,* who is ideally self-controlled (in Mele's sense of such control (1995, p. 121)) is ultimately a pawn of his brainwasher (p. 122); Condition 2 can be satisfied despite *S*'s beliefs being contrived by his deceiver (p. 180); and Condition 3 can be fulfilled in spite of *S*'s deliberative skills, habits or dispositions being engineered into *S* by *S*'s neurosurgeon (p. 184). So mind control does not undermine self-control, deception does not necessarily thwart autonomy, and engineering need not block autonomous deliberation. Conditions 0, 2 and 3 appear, therefore, to be non-historical.

Regarding beliefs (Condition 2) and means/end deliberation (Condition 3), appeal can be made to criteria other than causal history, such as truth versus falsehood, reliability versus unreliability, and rationality versus irrationality, to distinguish agents who are autonomous relative to their beliefs and deliberation from agents who are non-autonomous relative to these things. These other criteria are partly normative, whereas the criterion of causal origin and genesis is descriptive. As to failure in satisfying Conditions 2 and 3, Mele himself admits that historical, externalist criteria might be irrelevant:

> [S]omeone whose beliefs are not conducive to informed deliberation about some matters that concern him, *whatever the source of his doxastic state,* falls short of satisfying condition 2; and an unreliable deliberator, *independently of the etiology of his unreliability,* fails to satisfy condition 3. (1995, p. 188)

It seems, then, that Condition 1 bears the primary weight in distinguishing autonomous motivational states from non-autonomous ones in Mele's history-sensitive externalism. Drawing on an analogy with the internalism/externalism debate in the philosophy of mind and language,[7] Mele writes,

> [The autonomous possession of a pro-attitude], requires [not] that the agent *have* a history of a certain kind, but rather that he *lack* a certain kind of history—a history yielding what I have called "compulsion*" of [that attitude] . . . Hilary Putnam argued, famously, that "meanings just ain't in the *head*." If I am right, psychological autonomy ain't in the head either; or rather, it ain't *all* in the head. There is also a negative historical constraint on the autonomous possession of pro-attitudes: what I have called "authenticity" (1995, pp. 172–73).

Although the compatibilist trio, together with self-control, provides only sufficient conditions for autonomy, we interpret this (negative) historical, authenticity constraint as *necessary* for autonomy.

To formulate this constraint, we note, first, that Mele distinguishes between "being compelled to *acquire* a value [or pro-attitude] at a time and being compelled to *possess* a value [or pro-attitude] over a stretch of time" (1995, pp. 158–59). A value that an agent is compelled to acquire may or may not be sheddable, but "an agent's being practically unable to abandon during *t* a pro-attitude [or value] of which he is possessed throughout *t* is a necessary—but not a sufficient—condition of his being compelled to possess that pro-attitude [or value] (over that interval)" (1995, p. 166). The negative historical constraint that, it appears, Mele does accept is this:

> *Bypass-M:* An agent autonomously possesses an unsheddable pro-attitude *P* throughout an interval *t,* only if he is not compelled* to possess *P* during any segment of *t*.

Mele elaborates:

> 1.' A necessary condition of an agent *S*'s *authentically* possessing a pro-attitude *P* (e.g., a value or preference) that he has over an interval *t* is that it be false that *S*'s having *P* over that interval is, as I will say, *compelled**—where compulsion* is compulsion *not arranged by S*. (1995, p. 166)

As an approximation of a sufficient condition for *P*-compulsion*, Mele offers this (1995, p. 172):

> 1*. If an agent *S* comes to possess a pro-attitude *P* in a way that [i] bypasses *S*'s (perhaps relatively modest) capacities for control over his mental life; and the bypassing issues in [ii] *S* being practically unable

> to shed *P;* and the bypassing was [iii] not itself arranged (or performed) by *S;* and [iv] *S* neither presently possesses nor earlier possessed pro-attitudes that would support his identifying with *P,* with the exception of pro-attitudes that are themselves practically unsheddable products of unsolicited bypassing; then *S* is compelled* to possess *P.*

Brief commentary on elements [i] to [iv] will be useful. [i] *S*'s control capacities are bypassed in *P*'s generation if *P* is acquired in the absence of these capacities being operative or exercised. Ordinary people possess the basic control-capacities of ideally self-controlled agents in some measure:

> Such [ideally self-controlled] agents are [a] capable of modifying the strengths of their desires in the service of their practical, evaluative judgments. . . . They are [b] capable, moreover, of rationally assessing and revising their values and principles, of identifying with values of theirs on the basis of informed, critical reflection, and [c] of intentionally fostering new values and pro-attitudes in themselves in accordance with their considered evaluative judgments. (1995, pp. 166–67)

Control capacities are, in short, capacities to change a pro-attitude's strength, or to revise or eradicate a pro-attitude, or to acquire a pro-attitude hitherto unpossessed on the basis of critical reflection and evaluative judgment.

[ii] *S* is practically unable to shed pro-attitude *P,* if under ordinary circumstances *S* is unable either to eradicate *P* or to attenuate *P*'s strength (even though *S* would shed *P* under certain exceptional or counterfactual circumstances) (1995, pp. 153–54). [iii] *P*-compulsion* is not self-induced by *S,* in contradistinction to the self-arranged strategy of Ulysses to bind himself to the mast in order to resist the sirens' call and thereby to retain his autonomy (Elster 1984, pp. 36–47). [iv] *P*-compulsion* excludes self-identification with *P* unless the identification itself is explained by a bypassing of *S*'s control-capacities (1995, p. 171).

Armed with *Bypass-M,* Mele proposes a solution to the problem of CNC manipulation. Against the incompatibilist, he holds that even at a causally deterministic world there is a marked distinction between a CNC manipulated victim—a compelled* subject ("*S**" for short)—and an autonomous agent *S,* precisely because the unsheddable pro-attitudes of the former are compelled* whereas those of the latter are authentic. Against the internalist compatibilist, Mele holds that CNC mind-control destroys internal autonomy, specifically because CNC controlled unsheddable pro-attitudes are compelled* attitudes. So, although *S* and *S** are both causally determined and are psychological twins "from the inside," only *S* is autonomous because *S* lacks a compulsion* history.

Mele's account is a *significant* improvement over other historical, externalist accounts, for it enunciates an alluring principle instead of delivering a conventional, stipulative list in response to the question of what

distinguishes "illegitimate external influences" from legitimate ones.[8] Illegitimate external influences on pro-attitude, *P,* yield *P*-compulsion*. The authenticity condition (that *Bypass-M* expresses) tries to capture why different instances of CNC manipulation or mind control, such as secret hypnosis, subliminal advertising, clandestine electronic brain stimulation, unperceived brainwashing, concealed conditioning, and so forth, are all illegitimate (when so) in the sense of being autonomy undermining.

The principled distinction *Bypass-M* draws between compulsion* and noncompulsion*, according to Mele (1995, p. 158), resorts to a familiar compatibilist distinction which goes back to Hobbes (1651/1985, pt. 2, chap. 21) and Hume (1739/1975, bk. 2, pt. 3, sec. 2), namely that between compelled (or constrained) and ordinarily caused behavior.[9] Although both causal compulsion and causation are deterministic or necessitating, only the first is inimical to autonomy. Causal necessitation as a result of compulsion is autonomy-thwarting, whereas "mere" deterministic event causation is compatible with autonomy.

We are in sympathy with, and applaud, most of these penetrating insights of Mele. We confine attention to what appears to be a troubling consequence of the "Bypass View": *Bypass-M* (together with an empirical assumption to be introduced) seems to imply that authentic education is, largely, unfeasible.[10] Discussing various complexities of his history-sensitive authenticity condition, Mele develops the following "extreme" case of "manipulative indoctrination":

> A religious fanatic initially conducts his child's religious instruction in the same matter-of-fact tone in which he teaches the child about mundane matters: there are birds, bees, bicycles, and buildings, and there is God, a devil, heaven, and hell. Then he teaches the child that there are a few extraordinarily evil people who "reject" God—the dreaded atheists—and that they will burn in hell forever, a prospect he graphically portrays. For good measure, he "informs" the child that people who even *entertain doubts* about his religious teachings also burn eternally. Suppose that the man does a thoroughly effective job, and the child—owing primarily to a deep-seated fear of eternal damnation—grows up with firmly held religious convictions and values that, even as an adult, he is practically unable to shed. Suppose further that very young children have no capacity for controlling what religious doctrines and values they accept, no capacity for making up their own minds about such matters. Then, it might be said, the father has compelled the child to have certain pro-attitudes without *bypassing* the child's capacities for control over his mental life: one cannot bypass nonexistent capacities. (1995, p. 167)

We do not dispute Mele's observation that "[e]ven young children—five-year-olds, say—sometimes believe and desire on the basis of an assessment of evidence concerning matters that they comprehend (e.g., after attending their

first concert, art exhibit, or circus, they may, on the basis of an assessment of their experience, believe that they would enjoy going to another one and desire to do so)" and, consequently, we agree with Mele's conclusion that, in the case of five-year-olds, "[t]he capacity so to believe and desire is *inoperative* in the imagined inculcation [of religious convictions and values]: it is circumvented" (1995, p. 167).

However, even "a directly relevant (modest) capacity to believe and desire on the basis of assessment of evidence" (Mele 1995, p. 167) has to emerge at a particular time, time *t*. Distinguish between two stages in a child's development: the stage prior to which the child has acquired an initial (modest) capacity for making up his own mind—the *pre*-initial "reflective control" stage (before *t*)—and the stage after acquiring such a capacity—the *post*-initial reflective control stage (after *t*). Let's say that a child is in *early infancy* at times at which she exists before *t*. While a capacity for deliberative control over beliefs and desires is, to some modest degree, present in the early post-initial reflective control stage, it is absent in the pre-initial stage. Hence, during the formation of pro-attitudes in early infancy, the child's control capacities are inoperative, not because they are circumvented but because very young children have no capacity for making up their own minds about any issue whatsoever; during the pre-initial reflective control stage such capacities are nonexistent.[11]

It would then seem that instilling pro-attitudes in early education "bypasses" the very young child's capacities for reflective control because these capacities are entirely absent at this stage. Add to this the empirical assumption that a significant cluster of the pro-attitudes acquired at the stage of early infancy are frequently unsheddable, and *Bypass-M* sustains a disturbing result: possession of such pro-attitudes are compelled*. The protest that one cannot bypass nonexistent capacities is beside the point. Imagine that as a result of early "training," Youngster acquires an unsheddable pro-attitude to avoid the company of people with dark skin. Pertinent behavior deriving (partly but principally) from this pro-attitude will not be behavior for which Youngster is morally responsible because Youngster will lack responsibility-grounding control over this behavior. A case of this sort may involve a Brandtian-type epistemic error—perhaps the error of having false beliefs about instrumental consequences. But coherent scenarios free of any such errors can be imagined in which the child ends up with a cluster of unsheddable pro-attitudes that give rise to later behavior for which she will not be responsible. With such cases, during the pre-initial reflective control stage of training, the child's control-capacities are "bypassed" in that the child cannot acquire (and possess) pro-attitudes without these attitudes failing to bypass the pertinent capacities, again because the child lacks these capacities.[12]

What of the empirical assumption though, the assumption that a sizeable number of the pro-attitudes possessed at the stage of early infancy are frequently unsheddable? What if it is false? Well, if false, it is contingently

so. We can imagine beings cognitively and psychologically very much like us save for the fact that most (or all) of their pro-attitudes acquired and subsequently retained at the stage of early infancy are unsheddable at the time of acquisition and then after. An authentic education for such beings would be ruled out.

In sum, the Bypass View (together with the empirical assumption), seems to imply that not only "indoctrinative" religious education but also "liberal" education in early infancy may well issue in possession of compelled*, inauthentic, pro-attitudes. Given the empirical assumption, since much of education in the pre-initial reflective control stage, at least regarding instilment of pro-attitudes, is inauthentic, in significant measure an authentic education is not in the cards. This consequence of Mele's Bypass View is troubling because, intuitively, much more of education is authentic (or so we shall argue), even in the earliest stages of the child's development, than the Bypass View appears to allow.

The views we have considered in this appendix provide valuable insights on varied aspects of the manipulation problem. Learning from these views, we have, as the reader should recognize, modified and supplemented some of them so that the resulting package is, in part, our forward-looking, relational account of authenticity.

APPENDIX B

A Hard-Line Reply to the Four-Case Argument

Recently, McKenna developed an interesting response to Pereboom's four-case argument (Chapter 3, Section 3.5). If cogent, his response sheds doubt on our rejoinder to Pereboom (Section 3.6). In this appendix, we assess McKenna's engaging response.

B.1. McKenna's Hard-Line Response

Questioning the four-sequence strategy, McKenna starts with a case in which an agent is covertly manipulated in some manner ("manner X," he says) into satisfying "all of the conditions sufficient for the Compatibilist-Friendly Agential Structure (CAS). CAS is meant by compatibilists to exhaust the freedom relevant condition for moral responsibility." (McKenna forthcoming, sec. 1) He then suggests that the incompatibilist, with her sights on a particular instantiation of CAS, advances the following argument which he dubs the *Manipulation Argument* (MA):

1. If S is manipulated in manner X to A, then S does not A of her own free will and is therefore not morally responsible for A'ing.
2. An agent manipulated in manner X to A is no different in any relevant respect from any normally functioning agent determined to do A from CAS.

3. Therefore, if S is a normally functioning agent determined to A from CAS, she does not A of her own free will and therefore is not morally responsible for A'ing. (McKenna forthcoming, sec. 1)

McKenna calls a compatibilist reply that rejects Premise 1 a *hard-line reply* and one that rejects Premise 2 a *soft-line reply*. We consider, first, McKenna's defense of the view that any soft-line reply will fail, and, second, his hard-line reply.

At the heart of McKenna's argument for the provocative claim that any soft-line reply is ultimately unsuccessful is the following:

> Given that it is a formal condition of compatibilism that CAS could arise from a determined world, I can see no way to foreclose the metaphysical possibility that the causes figuring in the creation of a determined morally responsible agent could not be artificially fabricated. . . . If so, a soft-line reply to a well-crafted version of MA can only temporarily forestall the inevitable. Let the compatibilist adopt the soft-line by resisting case after case, showing how in each it falls short of CAS. The troubling point for the compatibilist inclined to avoid the hard-line reply is that some credible manipulation case could be fashioned. (McKenna forthcoming, sec. 2, note omitted)

This line of reasoning, however, is not beyond suspicion. To expose one of its shortcomings, we recast the argument in a fashion that invokes what should by now be the familiar distinction between normal and deviant etiologies: some causal routes that culminate in action are *normal* in that a normal etiology—a normal causal history—does not compromise the freedom of an action. *Deviant* causal routes—or deviant causal histories—do undermine the freedom of actions.

As we have stressed, if a causal route that culminates in action, presumably via the agent's acquisition of salient action-producing elements, such as desires or beliefs, is deviant, it is deviant relative to a causal route that is normal. It is noteworthy that McKenna does not believe that all forms of manipulation undermine responsibility; his sensible view is that some do but that others do not. Hence, it appears that he should accept the proposal that some causal routes that terminate in action are deviant but others are normal.

Since any proponent of the MA or the four-sequence argument agrees (or should agree) that whereas some forms of manipulation undermine responsibility (the causal histories involving such manipulation are deviant), other forms do not (the causal histories involving these other varieties of manipulation are normal), and since it is crucial to the success of these arguments that the manipulation to which these arguments appeal *is* responsibility undermining, these arguments rest on the reasonable assumption that not all causal routes to action are deviant and that not all such routes are normal.

A *compatibilist-friendly causal route* is, roughly, a causal route to action that a compatibilist can accept as grounding free action and responsibility. So, for instance, Fischer and Ravizza (1998, p. 86) would say that if your action issues from a moderately reasons-responsive mechanism for which you have taken responsibility, then this is sufficient (assuming other conditions of responsibility are met) for your being responsible for this action. For Fischer and Ravizza, a causal route to an action of yours that issues from an "owned" moderately reasons-responsive mechanism is compatibilist friendly.

McKenna's *objection* to the view that any soft-line reply to the MA ultimately fails requires presupposing that any compatibilist-friendly causal route is deviant. Why so? Well, he submits that *no matter what* compatibilist candidate is on the table—Fischer and Ravizza's (1998), Frankfurt's (1971), etc.—"some credible manipulation case could be fashioned" (McKenna forthcoming, sec. 2) that calls this contender into question. The star character of the fabricated case can always be manipulated in such a way that, despite the manipulation, she exercises the sort of control that the compatibilist candidate in question demands for responsibility. Presumably, to be *credible,* the fabricated case must be one in which a manipulated individual, who satisfies the conditions for free action laid down by this compatibilist contender, is at least, on the face of it, *not* morally responsible for an action that she is manipulated into performing. If it were not reasonably clear that the manipulated victim was *not* morally responsible, there would be little reason to accept the claim, advanced by the proponent of the MA, that the agent is not responsible. So what is supposedly a compatibilist-friendly causal route that culminates in the pertinent action really qualifies as deviant—it undermines responsibility.

However if any compatibilist-friendly route is deviant, then it seems that any incompatibilist-friendly route is deviant as well. A properly crafted global manipulation case that incorporates nondeterministic or agent causation should leave unaffected the verdict that the victim of manipulation is not responsible for actions that express her engineered-in pro-attitudes, values, and other things. Thus, we believe that the premise, if every compatibilist route is deviant, then every route, compatibilist friendly or otherwise, is also deviant is plausible. It follows that there is no normal causal route, contrary to what must be presupposed, if McKenna's objection to the soft-line reply is to be cogent.[13]

It may be rejoined that McKenna's argument for the view that any soft-line reply is ineffective presupposes only that every compatibilist-friendly causal route *can* be such that causes figuring in the creation of a determined morally responsible agent could be artificially fabricated and *not* that any such compatibilist route *is* deviant. In any event, a case that presupposes the latter would merely beg the question against the hard-liner. For according to McKenna, the compatibilist's best strategy to oppose the four-case argument is to show how *similar* a determined agent is, for instance, to a globally manipulated one (McKenna 2005*b:* p. 217). The manipulated

victim is prima facie not morally responsible, but reflection reveals that the actions issuing from an appropriately manipulated agent should be evaluated no differently than the actions issuing from a possibly naturally determined agent. The whole point of the hard-line reply is exactly to oppose the view that the manipulated agent is not responsible. In short, if a compatibilist believes that an ordinary agent—a naturally determined one—who satisfies what this compatibilist regards as necessary and sufficient for freedom-level control *is* morally responsible for relevant actions (and so the compatibilist causal route to these actions is *not* deviant), then why should this sort of compatibilist not also embrace the view that an agent who is the victim of manipulation *but who satisfies these very conditions* (despite the manipulation) is also responsible? This is the thunder of the hard-liner.

To meet this rejoinder, we backtrack for a bit: a soft-liner rejects the premise that an agent who does *A* as a result of manipulation does not differ in any relevant respect from a normally functioning agent whose *A*-ing is both causally determined and respects the strictures of freedom of the compatibilist contender in question. The rationale for McKenna's interesting submission that any soft-line reply will, in the end, fail is that some credible manipulation case can always be advanced to tell against the compatibilist contender. One merely needs to be imaginative enough to fabricate the undermining cases. Simply make sure that, despite being manipulated into performing some action, the relevant character in the fabricated case satisfies the freedom requirements of the compatibilist contender under scrutiny.

However, as we see things, what bears emphasis is that this sort of reasoning to question the soft-liner's response *presupposes* that the manipulation featured in the cooked up case *is* responsibility undermining and not just that it *could* be responsibility undermining. To elaborate, contrary to McKenna, in our assessment, to be a case in which it is *credible* to suppose that the manipulated agent is *not* responsible in virtue of being manipulated, the fabricated case involving manipulation, that is designed to undermine the compatibilist contender under assessment, must presuppose that the compatibilist route to *A*-ing—the causal trajectory of *A* (*A* is the action that the agent like Plum or Beth is manipulated into performing) that satisfies the freedom conditions of the compatibilist contender—*is* deviant. For if the compatibilist route to *A*-ing were *not* deviant—if it were, instead, normal or if it could be normal—then *contrary to what McKenna's argument against the soft-line response itself implies,* the sort of manipulation called upon to impugn the compatibilist contender would *not* undermine freedom. It would not do so because there would be no freedom to *be* undermined in the first place, on the assumption that the compatibilist route to *A*-ing were normal or could be normal! Consequently, it would not be true that the newly fabricated case would cast suspicion on the compatibilist contender. This, in turn, would spell victory for the soft-liner.

Further, we do not see how the presupposition—*PS*—that, *to be credible, the newly fabricated case be one in which the compatibilist-friendly causal route leading to the manipulated individual's doing* A *be deviant* begs the question against the hard-liner. The soft-liner uncovers a presupposition (*PS*) of what may be a *hard-liner's* argument that any soft-line response ultimately fails. (The advocate of this argument need *not,* of course, be a hard-liner; he or she might simply be an interested party who has no prior commitment to being either a hard-liner or a soft-liner.) This hard-liner's argument features what we may refer to as the "hard-liner's newly fabricated case" that is meant to undermine the compatibilist contender. Suppose this case is some variation of a global manipulation case in which the agent *A*-s in virtue of being manipulated. According to presupposition *PS,* the agent's compatibilist route to *A*-ing is deviant. We are entertaining the objection that this presupposition begs the question against a hard-liner such as McKenna. Now it is quite correct to point out that this implication of presupposition *PS*—that the agent's compatibilist route to *A*-ing is deviant—runs contrary to the hard-liner's proposal that a suitably determined agent who satisfies freedom conditions of the targeted compatibilist contender is pertinently similar to the globally manipulated agent; the hard-liner wants to say that the manipulated agent's causal route to *A*-ing, in the fabricated test case, is *not* deviant. Nevertheless, we fail to see how this implication of the presupposition begs the question against the hard-liner: if the hard-liner's argument against the soft-liner is to succeed, presupposition *PS,* as we explained, must be true. This presupposition or an implication of it may well be contrary to *other* claims of the hard-liner, such as the claim that the manipulated agent (in the fabricated test case) *is* responsible. But if it is so, its being so reveals a tension between presupposition *PS* and these other claims. This is a far cry from *begging* any questions against the hard-liner.

It goes without saying, we hope, that our account of relational authenticity helps to underpin a soft-line reply that arguably escapes McKenna's concern about any such reply. We do not, for example, see how a manipulation case can be constructed in such a way that the manipulators both engineer into, say, Beth, desires and beliefs in a manner that leaves Beth's evaluative scheme on the sidelines and, at the same time, the manipulation engages elements of Beth's scheme.

Irrespective of whether one sides with McKenna's argument for the conclusion that any soft-line reply ultimately fails, the soul of McKenna's engaging hard-line reply merits careful scrutiny. McKenna insists that we should help make the first two cases involving manipulation better by embellishing the cases, should there be need to do so, in a fashion in which it becomes pellucid that the victims of manipulation satisfy the proposed compatibilist sufficient conditions of freedom in question. The more confident we are that the manipulated agent does satisfy these conditions, the more the compatibilist who is a hard-liner would be inclined to hold that the agent *is* responsible despite the manipulation. McKenna writes,

> I propose a four-step [hard-line] reply to any instance of MA. *Step One: Reject all non-starters*. Consider the example. See if it is in the running for CAS. If not, the jig is up. Reject premise 2 and be done. *Step Two: Help make the manipulation cases better*. If the example gets past step one, if it comes close to getting CAS right but falls shy, amend the example. Help out your "good friend" the incompatibilist so that the example does get CAS right. This calls into relief that manipulation can be "just like" determinism. *Step Three: Fix attention on salient agential and moral properties*. Illustrate how the agent manipulated in manner X to satisfy CAS lives up to a rich sort of agency and genuinely satisfies certain moral properties (for example, does moral wrong). *Step four: Make clear that "manipulation" is not all that uncommon*. Lessen the intuitive uneasiness of the claim that an agent manipulated in manner X is free and responsible by calling attention to mundane causal factors that have a similar result, but are not thought to be freedom or responsibility undermining. (McKenna forthcoming, sec. 2)

B.2. Problems With the Hard-Line Reply

We believe though, that there are concerns with this way of interpreting the four-case strategy; it is not dialectically the most charitable way to understand Pereboom's argument. We are troubled, in particular, with McKenna's second step, the "embellishing" or "bolstering" step. As McKenna is well aware, the four-case argument is selective in that it proceeds by targeting, severally, specific compatibilist (or libertarian) candidates. Limiting attention to compatibilist candidates, there is no consensus among compatibilists regarding when manipulation is menacing—when it is freedom or responsibility subversive—and when it is benign. Reconsider Pereboom's Case 2 which may well be an instance of global manipulation akin to the Beth/Manson scenario or Psychohacker. Free will theorists differ over whether the agent in question (Plum or post-surgery Beth, for instance) is morally responsible for actions that express the agent's engineered-in pro-attitudes, values, deliberative principles, and so forth. Haji (1998), Kane (1996), and Mele (1995; 2006), for instance, think that the agent is not morally responsible for these actions; McKenna and Frankfurt believe otherwise.

We submit that if the generalization strategy is to be regarded as even prima facie tenable, the first two cases involving manipulation in the four-sequence progression must command the allegiance of targeted compatibilists and libertarians: the targeted audience must concur that the manipulation in question is of the variety that, on the face of it, *does* threaten free action or responsibility. For if the manipulation were of the sort that a targeted party deemed not to be of concern, *ab initio* the four-case sequence would have *no purchase at all* on this party. If anything is clear, the literature reveals that targeted compatibilists have, generally, not taken the four-case argument to be toothless, something that would

be difficult to appreciate if the first two cases were to be bolstered in the way in which McKenna suggests. To bring out this point, contrast two approaches to the four-case argument by compatibilists who defend different conditions on free action.

Imagine, first, that the four-case argument has its sights on Frankfurt and take Pereboom's Case 2 to be a case in which Plum is globally manipulated; he is subject to a similar sort of manipulation as Beth is. Frankfurt has persisted in maintaining that as long as the agent's action non-deviantly arises from a first-order desire with which the agent identifies, the agent is responsible for the action no matter what the provenance of the agent's psychological repertoire:

> A manipulator may succeed, through his interventions, in providing a person not merely with particular feelings and thoughts but with a new character. That person is then morally responsible for the choices and the conduct to which having this character leads. We are inevitably fashioned and sustained, after all, by circumstances over which we have no control. The causes to which we are subject may also change us radically, without thereby bringing it about that we are not morally responsible agents. It is irrelevant whether those causes are operating by virtue of the natural forces that shape our environment or whether they operate through the deliberate manipulative designs of other human agents. (Frankfurt 2002, pp. 27–28)

> [T]o the extent that a person identifies with the springs of his actions, he takes responsibility for those actions and acquires moral responsibility for them; moreover, the questions of how the actions and his identifications with their springs are caused are irrelevant to the questions of whether he performs them freely or is morally responsible for performing them. (Frankfurt 1975/1988, p. 54)

In Frankfurt's assessment, it would seem, the manipulation in the first two cases is benign.[14] Regarding responsibility, Frankfurt (2002, pp. 27–28) affirms that it is immaterial what the sources of our desires are—it doesn't matter whether our desires are caused by the "natural forces that shape our environment" or whether "they operate through the deliberate manipulative designs of other human agents"; we are responsible for actions that causally issue from our first-order desires as long as (other conditions of responsibility satisfied) we identify with these desires. In the proposed case, Plum is manipulated into performing some action but the manipulation leaves intact the sort of control that Frankfurt demands for responsibility: Plum identifies with the first-order desire from which the action that he is manipulated into performing causally arises. Frankfurt would simply regard this sort of manipulation as benign; so he wouldn't take this sort of test case to undermine his compatibilist account of control.

Imagine, next, that the four-case argument is launched against Mele. We have taken note that, unlike Frankfurt (1975/1988; 2002), Mele believes that globally manipulated agents are not morally responsible for their germane actions (see Appendix A, Section A.4). It is, thus, not surprising that, with certain amendments to Case 2, Mele accepts the judgment that Plum is not morally responsible for killing White in this case. He argues though, that this judgment can be endorsed consistently with maintaining that Plum may well be responsible for his murderous deed in Case 4. In other words, Mele argues for a soft-line reply. Mele as we previously documented, theorizes that if various historical considerations regarding the acquisition of actional antecedents, such as desires or beliefs, are not met, the agent is not responsible for behavior to which these antecedents give rise. If, for instance, an agent acquires a pro-attitude such as a desire via a process that totally bypasses the agent's normal capacities of deliberative control, the agent is practically unable to shed this pro-attitude, and the bypassing was not itself arranged (or performed) by the agent, then the agent is not responsible for acquiring that pro-attitude and is not responsible for actions that express that pro-attitude (Mele 1995, pp. 171–22; 2006, p. 170) We may say that a bypassing condition on free action is a condition that is not satisfied when an agent acquires a pro-attitude (or other actional elements) in a manner described in the last sentence. Assuming Mele is the quarry, if we were to try to amend Case 2 (and Case 1) so that all the historical conditions including the bypassing condition that Mele deems relevant to free action were satisfied, Case 2 would no longer qualify as a case of global manipulation. It would then, once again, be puzzling why Mele should expend any energy on the four-case argument if the first two cases were to appeal to manipulation that Mele regards as *benign*.[15]

Reverting to the supposition that it is Frankfurt's theory (1971/1988) that is the prey of the four-case argument, would this argument not minimally show that since Frankfurt's account of freedom generates the result that the manipulation in the first two cases is benign—the account generates the result that manipulated Plum and Beth *are* responsible for their actions despite the manipulation—the account, is therefore mistaken? In other words, we are to imagine someone as responding to Frankfurt in this way: "In the fabricated test case against your conception of control, Plum is manipulated into performing the action in question—killing White. What does it matter that Plum identifies with the first-order desire that causes him to kill White? *Because* he kills as a result of *manipulation,* Plum is *not* responsible for this murderous deed. If your account of control implies otherwise, so much the worse for the account." However, the four-case argument, *on its own,* does not support this sort of retort to Frankfurt. This is because when the argument is targeted at Frankfurt's compatibilism, its first two cases *assume* that Plum is not morally responsible despite Plum's satisfying Frankfurt's conditions on free action. That is, the first two cases simply assume, contrary to Frankfurt (1975/1988; 2002), that

the manipulation is *not* benign. So we have this sort of dialectical situation: Frankfurt insists that Plum, despite the manipulation, *is* morally responsible for killing White because Plum identifies with the relevant first-order desires; the proponent of the four-case argument *denies* that Plum is morally responsible for the killing because Plum is manipulated into performing this action. Digging in one's heals in this way is not the way of progress. What is required to challenge Frankfurt are *independent* reasons that call into question his account of freedom or control or his account of when manipulation is benign rather than menacing.

Now we have a puzzle. What we have just said of how Frankfurt should react to the four-case argument is equally true of how other targeted compatibilists should react to this argument if the first two cases are *bolstered to accommodate their conditions on freedom*—if it is made plain that the manipulated agent in the fabricated test cases, despite the manipulation, satisfy the conditions of control of the compatibilist accounts in question: the proponents of these accounts, like Frankfurt, should first simply deny that the manipulation in the initial two cases is menacing and should then indicate that the four-case argument itself cannot impugn their account of freedom or their account of when manipulation is menacing and when benign. If the four-case argument can be stopped in its tracks, in this fashion, at the first two cases in connection with *any* compatibilist account, where then is its bite?[16] Why do compatibilists have anything to fear from this argument?

Compatibilists and libertarians agree that whereas some forms of manipulation are responsibility-subversive others are not. (They may disagree, of course, on which forms are benign and which menacing.) On the presumption that all cases of manipulation are not equal—they do not all subvert freedom or responsibility—these theorists must have some basis for distinguishing the benign cases from the menacing ones. Assume that there is a set of conditions, which we simply label "Benign-M," which is such that if an agent satisfies all the members of this set despite being manipulated, free agency is not compromised. And assume, further, that any well worked out compatibilist or libertarian account of free action includes, as a component, its candidate of what Benign-M is. We suggest that the undeniable allure of Pereboom's four-step argument is that, among interested parties in the free will debate, there is no consensus on what Benign-M is and, thus, it is not transparent at the outset whether Plum in the first two cases involving manipulation does satisfy Benign-M. Hence, it is prima facie credible that Plum, in these cases, may well not be morally responsible.

Tying some ends together; we suggest that it is *misleading* to construe the four-case argument as unfolding in this fashion: let Best-Theory refer to a targeted compatibilist or libertarian theory that has as a constituent its contender of Benign-M, and imagine that Plum in the first two cases satisfies the conditions of this contender. Should Plum fail to satisfy these conditions, embellish the case so that it is obvious that he does satisfy these

conditions (this is McKenna's bolstering step). Then the progression from Case 1 to Case 4 impugns Benign-M and, thus, Best-Theory. The easy (and astute) response to this way of interpreting the argument would be McKenna's hard-line response: the compatibilist (or libertarian) at issue should simply reject the claim that Plum in the first two cases is not responsible for his germane actions despite being manipulated. Rather, we suggest that to appreciate the dialectical pull of the four-case argument, Pereboom be taken to be saying something of this sort: "You, the compatibilist, agree that there are manipulation cases in which the agent, Plum, is *not* responsible *because* he is manipulated. After all, all you compatibilists agree that some cases of manipulation are cases in which the manipulation is benign whereas others are cases in which the manipulation is menacing. Understand the first two cases in the four-sequence argument as featuring manipulation which, given your compatibilist account of control, you think is menacing. Then I don't see how you can claim that Plum is also responsible in the fourth case. Since this is so, your compatibilist account goes down the drain. But what's true of your account is true of *any* plausible compatibilist contender, a contender that should imply that Plum is not responsible in the carefully-selected first two cases. So any compatibilist account fails."

Understood in this way, the four-sequence argument prompts the following challenge: what *is* the relevant account of manipulation (or free action) that generates the prima facie plausible verdict that Plum is not morally responsible in the first two cases? If we can uncover this account, we will subsequently be in a position to ascertain whether Plum in the fourth case is also not responsible. We propose (whether or not this was his original intention) to take Pereboom to be recommending that no libertarian account divorced from agent causation or that no compatibilist account that takes seriously the view that Plum is not responsible in the first two cases delivers a contrary verdict in the fourth case. We venture that this recommendation should not be heeded, for we believe that our relational account of authenticity can deliver the goods.

Notes

NOTES TO CHAPTER 1

1. As we expose later (Chapter 3, Section 3.4), there is an interesting complication with this example.
2. Robert Kane (1996) and Derk Pereboom (2001) have lucidly articulated this problem. See also Fischer and Ravizza (1998) and Mele (1995).
3. See for example, Haji and Cuypers (2001), Mele (1995, 2006), and Pereboom (2001).
4. See for example, Archard (1993, 2003*a*), Feinberg (1980, 1986), Gutmann (1980), and Peters (1963/1974, 1973/1974).
5. For a classic formulation of this problem—the "paradox of moral education"—see Peters (1963/1974). See also Cuypers (2008).
6. See for example, Bailin and Siegel (2003), and Siegel (1997, 1988).
7. See for instance, Cuypers (2004*a*), Dearden (1972), and Peters (1973/1974).
8. Pereboom (2001, 2002). Others of this bent include Richard Double (1991, 2004), Ted Honderich (1993), and Bruce Waller (1990).

NOTES TO CHAPTER 2

1. Aristotle, *Nicomachean Ethics,* 1109b30–35.
2. See for example, Haji (1998), Fischer and Ravizza (1998), and Mele (1995).
3. Further elaboration and defense of the first three conditions can be found in the literature on moral responsibility. On agency: see for example, P. Strawson (1962); on the epistemic requirement: see for example, Ginet (2000), Haji (1998), and M. Zimmerman (2004); and on control: see for example, Frankfurt (1969/1988, 1971/1988), Mele (1995), and Fischer and Ravizza (1998).
4. The authenticity condition should not be conflated with an alleged *autonomy* condition of responsibility. A slave, though non-autonomous, may well be morally praise- or blameworthy for many of her actions and intentional omissions. She would not be so if her behavior issued from causal springs that are inauthentic.
5. The manipulation problem is sometimes called the "source" problem. See for example, Fischer and Ravizza (1998, p. 200), and D. Zimmerman (2003*a,* pp. 639–40).
6. For elaboration, see Haji and Cuypers (2001).
7. Various authors have discussed such cases of global manipulation. See for example, Dennett (1984), Double (1989), Fischer and Ravizza (1998), Mele (1995, esp. ch. 9, 2006), and Haji (1998, esp. ch. 6).

8. Walter Glannon has argued that our practices of holding people morally and criminally responsible require only a low threshold of psychological connectedness and bodily continuity. See Glannon (1998, sec. IV, 2002, esp. ch. 4).
9. Mele (1995, p. 164). See also, Mele (2006, pp. 163–73).
10. See Fischer (1987), Fischer and Ravizza (1994), and Locke (1975).
11. See for example, Cuypers (2000*a,* 2004*b*), Fischer and Ravizza (1998, pp. 194–201), Mele (1995, pp. 59–85, 144–56), Shatz (1986), Slote (1980), Thalberg (1978), and Watson (1975).

NOTES TO CHAPTER 3

1. See for example, Watson (1999, pp. 360–65).
2. Non-causalists, such as Carl Ginet (1990) and Stewart Goetz (1998) will not accept this assumption.
3. See for example, Frankfurt (1971/1988).
4. Similarly, Alfred Mele claims that ultimate control involves the lack of deterministic causal influence upon one's action of agent-external events. For an agent to have ultimate control over, for instance, his making some decision, it should not be the case that there are minimally causally sufficient conditions, that do not include any event or state internal to the agent, for the agent's making this decision. Hence, agents could have ultimate control over their actions only if determinism is false. See Mele (1995, p. 211).
5. See for example, Clarke (1996, esp. p. 27); O'Connor (2002, esp. pp. 197–98).
6. Elsewhere, Haji (2001*a*) argued that the sort of indeterminism Kane's account of freedom assumes does not enhance the active or causal control the agent exercises in performing a free action in comparison to the control the agent wields over her actions on leading compatibilist accounts. Nor does this sort of indeterminism endow the agent with positive powers to influence which alternative, from a set of open alternatives, is made actual. (On this point, see Clarke 1996.) It should, then, come as no surprise that Kane's account of freedom succumbs to the problem of CNC manipulation. If one takes determinism to threaten responsibility because one regards determinism as relevantly analogous to CNC control, then, it seems, one should also take Kanian libertarianism to threaten responsibility for similar reasons.
7. Lynne Rudder Baker (2006) proposes that only agents with "first-person perspectives" can be morally responsible, and that this perspective cannot be acquired by neural manipulation. See also, Baker (1998).
8. See for instance, Clarke (2003, pp. 153–54); O'Connor (2000, pp. 52–53).
9. For elaboration, see Haji and Cuypers (2001, pp. 232–35).
10. Haji, for example, discusses the first two requirements, in his 1998, chs. 4, 8, and 9.
11. More fine-grained distinctions, such as the distinction between early childhood and adolescence, and the distinction between an inchoate and a mature evaluative scheme, are possible. The distinction we mark in the text suffices for our purposes.
12. There is a complication here that we sidestep: the acquisition of an evaluative scheme is a matter of degree; so, depending upon their stage of development, at various stages of maturation children may be partial normative agents.
13. See for example, Haji (2003*a*).
14. For concerns about this internalist (non-historical) suggestion, see Cuypers (2004*b*).
15. P. Strawson (1962) develops this connection between moral agency and the morally reactive attitudes.

16. Galen Strawson (1986, p. 293) suggests that the implantation or fostering of the attitude or disposition of seeing oneself in control of one's actions, as a suitable subject of moral responsibility, may well—at least in the very initial stages of development—be required to ensure responsibility for subsequent behavior. Fostering this sort of attitude is presumably morally required. For a similar view, see Fischer and Ravizza (1998, p. 208).
17. Here, we have drawn from Haji (1998, pp. 126–32).
18. Alfred Mele (2000) underscores the importance of neurophysiology in his response to the problem of causal deviance discussed in the literature on intentional action.
19. See for example, Kane (2000*b*); Pereboom (2001, pp. 110–17).
20. Our analysis of when an initial scheme is authentic implies that Fischer and Ravizza's third condition on taking responsibility, at least as formulated, seems not to be correct. Fischer and Ravizza (1998, pp. 210–14) explain that taking responsibility for a mechanism that issues in action has three components: (i) the agent must see himself as the source of his behavior; (ii) he must accept that he is a fair target of the reactive attitudes as a result of how he exercises his agency; and, (iii), the views of himself in (i) and (ii) must be based in an appropriate way on the evidence. Suppose, at the pre-initial scheme stage, neurologists implant in Youngster various pro-attitudes and beliefs, including the belief that Youngster sees herself as the initiator of her behavior and the belief that she regards herself as a suitable target of the reactive attitudes based on how she exercises her agency. Suppose that none of the implanted elements, including these engineered-in beliefs, subvert moral responsibility for actions issuing from the mechanism of practical reasoning and others that include these elements. Then this sort of manipulation is not responsibility-subversive. Fischer and Ravizza's third condition implies otherwise: it is not true that Youngster acquired relevant mechanisms in an appropriate way, namely, through response to moral training and acquiring knowledge of the world in normal ways.

NOTES TO CHAPTER 4

1. After "him" in the fourth line of the passage in the original, there is a reference to Frankfurt's *The Importance of What We Care About* (1988, p. 21).
2. See Mele (1995, p. 172).
3. See Mele (1995, p. 145).
4. Mele (1995, p. 175, note 22) suggests that in such transformation cases, the pre- and post-surgery agents may be strongly psychologically connected, in Parfit's sense. They may be such that the number of direct psychological connections between them "is *at least half* the number that hold, over every day, in the lives of nearly every actual person" (Parfit 1984, p. 206). In addition, Mele argues that the pre-surgery agent (t-Beth) just before her transformation is much more similar, on the whole, to the post-surgery agent (t*-Beth) than she is to neonate Beth or toddler Beth. Still, t-Beth is the same person as the neonate and toddler Beths, in a familiar "personal identity" sense of "same person." So what is to prevent her from being the same person, in the same sense, as t*-Beth?
5. For further diagnosis of Frankfurt's stance on such manipulation cases, see Cuypers (2004*b*).
6. To anticipate, the thought here is *something* like this: If the identification view is correct, then manipulated Beth is morally responsible for her germane actions; it's false that she is responsible for these actions. Therefore, the identification view is not correct.

7. On the coherence of instantaneous agency and the implications of such agency for the internalism/externalism debate, see David Zimmerman (1999).
8. See Chapter 3, Section 3.3.
9. Compare this with Mele's treatment of pertinent manipulation that we outlined in Appendix A, Section A.4.
10. For further elaboration, see Cuypers (2004*b*).
11. See Haji (1998); Fischer and Ravizza (1998); Mele (1995, 2006).
12. For an extensive discussion on the concepts of activity, passivity, and volitional necessity that are relevant to this objection, see Cuypers (2000*a*).
13. See Mele (2006, pp. 179–84).
14. See Arpaly (2003, pp. 126–28), for development of this sort of objection.

NOTES TO CHAPTER 5

1. The examples Winch and Gingell (1999) advance are illuminating.
2. Further elaboration on the relationship between autonomy and authenticity can be found in Cuypers and Bonnett (2003). See also Cuypers (2001, part II).
3. See also, Archard (1993, part I).
4. For further details about this second set of examples, and a list of references to the progressivism literature, see Darling and Nordenbo (2003). See also, Darling (1994).
5. Archard (2003*b*) has a revealing discussion on this third class of examples. See also, Archard (1993, part II & III); Archard (2003*a*). For more on the parent–child relationship, see Smeyers and Wringe (2003).
6. People who discuss the view that critical thinking is an educational ideal (see, this chapter, Section 5.3) or those engaged with Philosophy for Children (see, for instance, Murris 2000) advance other examples that appeal to the child's authenticity. Apparently, their examples pertain to what they claim is the authenticity of rational thought and not, for example, to the authenticity of pro-attitudes or choices.
7. For relevant incompatibilist literature, see for example, Kane (1996); Pereboom (2000). For relevant compatibilist literature, see for instance, Fischer and Ravizza (1998); Haji (1998); Mele (1995).
8. See Feinberg (1986, p. 34–35). See also, Feinberg (1980, pp. 148–51); Gutmann (1980).
9. White (1990, pp. 25–26) argues that there are positive grounds in favor of some non-autonomous conceptions of well-being. For more on the debate between liberals and non-liberals, see Mulhall and Swift (1996).
10. For further elaboration and criticisms of this hierarchical account, see Cuypers (2000*a*, 2004*b*).
11. See for example, Bailin and Siegel (2003); Siegel (1988, 1997, 2003).
12. See for instance, Cuypers (2004*a*); Dearden (1972); Peters (1963/1974, 1973/1974).
13. Siegel acknowledges that his view is that "autonomy is a *necessary but not sufficient condition* of critical thinking" (Siegel 2005, pp. 542–43). This is because the autonomous person's reasoned appraisal of candidate beliefs may be so deficient that the appraisal fails to satisfy the "epistemic quality" demands of the reason assessment component. For our concerns, it suffices that Siegel holds that autonomy is necessary for critical thinking—both the reason assessment and critical spirit dimension require autonomy.
14. Siegel (2003, p. 307).
15. Dearden (1972, p. 70).

16. Some of the more important pro-attitudes and dispositions required for critical thinking include "*respect for reasons and truth* (commitment to having justified beliefs, values and actions); . . . *an inquiring attitude* (inclination to assess the support for judgements one is asked to accept); *open-mindedness* . . . *fair-mindedness* . . . *independent-mindedness* (possession of the intellectual honesty and courage necessary for seeking out relevant evidence and basing one's beliefs and actions on it, despite pressures or temptations to do otherwise, and the personal strength to stand up for one's firmly grounded beliefs); . . ." (Bailin et al. 1999, pp. 294–95).
17. For Peters' articulation of the "paradox of moral education" and his response to it, see Peters (1963/1974).
18. This case is modeled after the one Mele advances in Mele (1993, p. 275).
19. Development of the notions of being autonomous relative to the acquisition of a pro-attitude, relative to the possession of a pro-attitude, and relative to the influence of a pro-attitude can be found in Mele (1993, pp. 275–77; 1995, pp. 138–39).
20. Ever since *Relativism Refuted* (Siegel 1987), this appeal to rationality's self-justification is central to Siegel's work on the theory of rationality and the foundations of critical thinking as an educational ideal. Formulations of this sort of "transcendental argument" may be found, for example, in Siegel (1988, pp. 74–76, 132; 1997, pp. 81–87; 1998, pp. 30–31).
21. See for example, Haji (1989); relevant papers in Campbell and Sowden (1985); the papers in Paul et al. (1988); and various articles in Vallentyne (1991).
22. For critical discussion of Christman's view, see Haji (1998, pp. 90–94); Mele (1993).
23. We remain neutral on whether the causation in question is deterministic or nondeterministic.

NOTES TO CHAPTER 6

1. See also, Pereboom (2001, ch. 5, esp. pp. 156–57).
2. See Haji (2002).
3. Arguments for the incompatibility of determinism and alternative possibilities have been advanced by, among others, Ginet (1983, 1990); J. Lamb (1993); Fischer (1994); van Inwagen (1983); Warfield (1996); Wiggens (1973).
4. See for example, Feldman (1986); M. J. Zimmerman (1996).
5. Each morally deontic act proposition has this form: As of time, *t*, agent *S* morally ought (or it is morally permissible for *S*, or it is morally wrong for *S*), to do action *A* at time *t**.
6. See for example, Murphy and Hampton (1988, p. 20).
7. See also, Oakley (1992, pp. 122–59).
8. See for instance, Hume (1748/1981, section VIII); P. Strawson (1962/1982, pp. 67–70).

NOTES TO CHAPTER 7

1. On this distinction, see Williams (1985, pp. 6, 174–96); also, Mackie (1977, pp. 106–07).
2. On varieties of normative responsibility, see Haji (1998, pp. 177–96).
3. Much of what we have to say on normative blameworthiness will also apply, with necessary amendments, to normative praiseworthiness.

4. Michael Slote (1983, p. 86) develops an interesting example in which a father deliberately does something he believes to be morally wrong—he misleads the police about his son's whereabouts—taking the verdict of parental love to do whatever he can to save his offspring, to override the verdict of morality.
5. Roger Lamb (1997) proposes that love involves being committed to the beloved where the sense of "commitment" is a sense referring to our obligations as lovers (p. 28).
6. See also, Foot (1978); Stocker (1990); Williams (1976*a*); Wolf (1982).
7. For a defense of this view, see Haji (1998, pp. 140–67).
8. Here we ignore worries of determinism.
9. See for instance, Michael Zimmerman (2001, pp. 2–3). According to Zimmerman, a person's values are those things that are valued by (not valuable for) that person.
10. O. H. Green (1997) argues that love is not an emotion but a complex conative state, a set of desires. Green, though, does not deny that there are reasons for love.
11. Indeed, Green's position is stronger. Regarding romantic love, he claims that *A* loves *B* if and only if *A* desires to share an association with *B* which typically includes a sexual dimension, *A* desires that *B* fare well for his or her own sake, and *A* desires that *B* reciprocate the desires for association and welfare. See Green (1997, p. 216).
12. See for example, Aristotle, *The Nicomachean Ethics,* (1966, Book IX, Chapter 5, 1166b30–1167a12; Book IV, Chapter 6, 1126b20ff); Oakley (1992, pp. 58–59).
13. Here is Williams' original example: "But this construction provides the agent with one thought too many: it might have been hoped by some (for instance, by his wife) that his motivating thought, fully spelled out, would be the thought that it was his wife, not that it was his wife and that in situations of this kind it is permissible to save one's wife" (Williams 1976*b,* p. 214).

NOTES TO CHAPTER 8

1. Michael Zimmerman (1997, pp. 235–36). See also, Michael Zimmerman (1988, esp., pp. 38–54). Other partisans of the view that praise- or blameworthiness requires action on the basis of suitable beliefs of right and wrong include Brandt (1958, pp. 38–39); Milo (1984, ch. 1).
2. Here, again, bear in mind the distinction to which we previously called attention between a narrow and a broad conception of morality.
3. Needless to say, we are assuming that an account of acting out of friendship is divorced from the relevant morally deontic beliefs.
4. An advocate of the first thesis is Henson (1979). Supporters of the second thesis include Beck (1960, p. 228); Paton (1964, p. 19); Wolff (1973, p. 66); Michael Zimmerman (1988, p. 51).
5. See Smith (1996, pp. 181–83). See also, Smith (1994, pp. 74–76). For additional comments on Smith on morally perfect agents, see Mele (2003, pp. 123–25).
6. An especially insightful paper on acting from virtue is Audi (1995).
7. See for example, Pettit (1997, pp. 154–55).
8. Michael Zimmerman (1996), for example, defends a non-Kantian analysis of the concept of duty.

NOTES TO CHAPTER 9

1. See also, Haji (1998, pp. 151–67).

2. This species of argument is carefully laid out by Gowans (1987, pp. 20–22); McConnell (1978, pp. 155–57). It is invoked by Lemmon (1962) to discard **OMC**.
3. See for example, Hudson (1986, pp. 16–18); van Fraassen (1973/1987, p. 146); Williams (1973, pp. 132–33).
4. For arguments in favor of such moral dilemmas, see for example, Sinnott-Armstrong (1988); Stocker (1990); Marcus (1980/1987); Williams (1973). For arguments against such dilemmas, see for instance, Conee (1982/1987); Donagan (1984/1987); Feldman (1986, sec. 9.1); McConnell (1978/1987); Michael Zimmerman (1996, pp. 217–25).
5. See also, Frankfurt's example of Agamemnon at Aulis as an instance of "[s]ituations in which it is impossible for a person to avoid this sort of self-betrayal [that] provide the theme for one variety of human tragedy" (Frankfurt 1994/1999, p. 139, n. 8).
6. For objections to ideals in ethics and politics, see for example, Berlin (1988/1992). For arguments in support of ideals in education, see for example, Emmet (1994); Huxley (1937); Rescher (1987); Scheffler (1971/1973). For a discussion of these objections and defences, see De Ruyter (2006, 2007). For a critique of De Ruyter's defence of educational ideals, see Heyting (2004).
7. Here we proceed with caution: the import of such phrases as "being intrinsically (or extrinsically) related to certain kinds of practice" is not transparent to us.
8. Elsewhere White treats autonomy as a "central liberal value" that is predicated on the "fundamental value" of personal well-being: "Personal autonomy is a central liberal value. It rests on an even more fundamental value in human life—personal well-being. Autonomous well-being is only one variant of the more general concept, given that people can flourish or not flourish in non-liberal—for example, traditional-tribal—as well as liberal societies" (White 1999, p. 193). See also, White (1982, 1990, 2003). Standish (1999, pp. 35–40) advances a clear and succinct description of Peters,' Hirst's, and Dearden's liberal position. Christopher Winch's distinction between "weak" and "strong autonomy" is also relevant to the debate surrounding the liberal conception of personal autonomy; see Winch (1999, 2002).
9. For elaboration, see for example, Copp (1997). On plain "ought," see for instance, Feldman (1986).
10. See for example, Copp (1997); Michael Zimmerman (2001, pp. 239–41).
11. For more on this version, see Haji (forthcoming (b)).
12. Michael Zimmerman (2001, pp. 195–98) proposes that attitudinal pleasures and displeasures do have an affective aspect, so an adequate account of the nature of attitudinal pleasure and displeasure must make reference to their affective aspect; and that an adequate account of the value of these attitudes must also make reference to this aspect.
13. Not everyone agrees that states of affairs are bearers of intrinsic value. For example, Noah Lemos (1994, pp. 23–25) describes what he takes to be the bearers of intrinsic value as abstract but it appears that he does not believe that these bearers are states of affairs. G. E. Moore at times, talks of individual physical objects, such as books, as having intrinsic value (Moore 1903, p. 3). Michael Zimmerman (2001, pp. 50–52) proposes that concrete events are the bearers of intrinsic value.
14. This axiological principle does not take into account whether the pleasure is taken in an object that deserves to have pleasure taken in it. One might, for example, insist that deserved pleasure in someone's pain or in someone's undeserved pain is intrinsically bad. (The other four principles too, ignore objectworthiness.)

15. For further discussion on the intrinsic value of worlds and lives, see Haji (2004; forthcoming (a)).
16. See also, Feldman (1992, pp. 201–02); Rescher (1966, pp. 73–83).
17. We believe that it is much more promising that love contributes to the value of worlds; world betterment should be an educational aim.
18. On this theme, see for example, McKenna (2005*b*); Gary Watson (1987*a*).

NOTES TO APPENDIX A

1. Recent discussion of these issues appears in a book symposium on *Fischer and Ravizza's Responsibility and Control* in *Philosophical Explorations* Vol. 8, Nr. 2 (June 2005), pp. 91–156.
2. See Chapter 6, Section 6.2.
3. This theme is elucidated further in Chapter 4, Section 4.3.
4. For criticisms of Wolf's reason view, see for example, Fischer and Ravizza (1998, pp. 55–61); Haji (1998, pp. 65–70).
5. The notion of *something's being an evaluative scheme* is elaborated in Chapter 3, Section 3.4.
6. In the literature various labels mark this in-house division among compatibilists. For example, "structural" versus "historical" (Frankfurt), "non-historical/current time-slice" versus "historical" (Fischer and Ravizza), and "internal" versus "external" (Mele). David Zimmerman (2003*a*) has some insightful things to say about what these labels might try to capture. In Zimmerman's terminology, Mele is a *negative* source-historicist (David Zimmerman 2003*a,* pp. 646–47).
7. Cuypers has an instructive discussion of the externalism/internalism or historical/non-historical distinction in the debate on personal autonomy in his 2000*b* and 2004*b*.
8. Christman (1987, pp. 287–92), for example, advances such a list.
9. See also, Ayer (1954).
10. Cuypers (2006) discusses what he takes to be other troubling consequences.
11. Mele (1995) claims that, even in the pre-initial relective control stage when the pertinent control-capacities have not yet emerged, "the very young" still "have the capacity to *develop* into individuals who would make up their minds" (italics added) and that, in the imagined case of religious indoctrination, "that capacity was bypassed—and, indeed, destroyed" (p. 168). However, as Kapitan (2000) critically observes,

    > [t]his . . . seems far too sweeping: perhaps at the time of our births, most of us *could* have developed a capacity to exert control over a wide variety of pro-attitudes even though we did not. In our early development each of us is subjected to physical and social forces of which we are largely ignorant, over which we have no control, yet from which we acquire values, beliefs, motivations, and capacities for rational evaluation that subsequently guide our choices and actions. These forces 'destroyed' any capacity to become a different sort of person with self-control regarding *any* unsheddable pro-attitude that we happen to have. Consequently, every unsheddable pro-attitude is compelled, and anyone with firm unshakeable principles of action ends up being inauthentic and non-autonomous. (p. 89)

12. The interested reader should note that Mele takes the position that, in some possible cases, agents come autonomously to possess even unsheddable values that are brainwashed in (Mele 1995, p. 171).

NOTES TO APPENDIX B

1. It will not help the objection to the soft line reply if it were proposed that any compatibilist-friendly causal route is prima facie deviant. (If a causal route is prima facie deviant, it is so relative to a causal route that is not prima facie deviant.) For then, it seems that any causal route is prima facie deviant. This would, consequently, violate the implicit presupposition of the objection that some causal route is not prima facie deviant.
2. For further diagnosis of Frankfurt's stance on such manipulation cases, see Cuypers (2004*b*).
3. We do not, of course, deny that some compatibilists, as Mele notes, may reject the four-case argument for reasons other than the reason that Case 2 is a case of global manipulation. Here is a revealing passage from Mele (2005), addressing Case 3, Mele writes,

   > When Plum grew older, was he able, on a compatibilist reading of 'able,' to alter his 'character'? More specifically, was he able—perhaps partly through reflection on his values and experiences—to make himself less egoistic and more sensitive to moral reasons or to act in ways that have this result? Pereboom does not say. If the rigorous training practices did not render Plum unable to do these things, and if he was able—in a compatibilist sense—to do them, typical compatibilists have no good reason to agree that Plum is not morally responsible for the killing. If, however, the manipulation was such as to render Plum unable to attenuate its effects, some compatibilists can agree that Plum is not morally responsible for the killing. (p. 79)

4. Interestingly, John Fischer (2004, p. 158) submits that Plum is responsible in the first two cases (though he is not blameworthy).

References

Adams, Robert M. 1999. *Finite and Infinite Goods: A Framework for Ethics*. Oxford: Oxford University Press.

Archard, David. 1993. *Children: Rights and Childhood*. London: Routledge.

Archard, David. 2003*a*. *Children, Family and the State*. Aldershot: Ashgate.

Archard, David. 2003*b*. "Children." In H. LaFollette, ed., *The Oxford Handbook of Practical Ethics*. Oxford: Oxford University Press (pp. 91–111).

Aristotle. *Nicomachean Ethics*. Translated by Richard McKeon. New York: Random House (1966).

Arpaly, Nomy. 2003. *Unprincipled Virtue: An Inquiry Into Moral Agency*. New York: Oxford University Press.

Audi, Robert. 1986. "Acting For Reasons," *The Philosophical Review* 95: 511–46.

Audi, Robert. 1995. "Acting From Virtue," *Mind* 104: 449–71.

Ayer, Alfred J. 1954. "Freedom and Necessity." In A. J. Ayer, *Philosophical Essays*. London: Macmillan (pp. 271–84).

Bailin, Sharon, Case, Roland, Coombs, Jerrold R., and Daniels, Leroi B. 1999. "Conceptualizing Critical Thinking," *Journal of Curriculum Studies* 31: 285–302.

Bailin, Sharon, and Siegel, Harvey. 2003. "Critical Thinking." In N. Blake, P. Smeyers, R. Smith, and P. Standish, eds., *The Blackwell Guide to the Philosophy of Education*. Oxford: Blackwell (pp. 181–93).

Baker, Lynne Rudder. 1998. "The First-Person Perspective: A Test for Naturalism," *American Philosophical Quarterly* 35: 327–48.

Baker, Lynne Rudder. 2006. "Moral Responsibility Without Libertarianism," *Noûs* 40: 307–30.

Beck, Lewis White. 1960. *A Commentary on Kant's Critique of Practical Reason*. Chicago: University of Chicago Press.

Benn, Stanley I. 1976. "Freedom, Autonomy and the Concept of a Person," *Proceedings of the Aristotelian Society* LXXVI: 109–30.

Berlin, Isaiah. 1988. "The Pursuit of the Ideal." In I. Berlin, ed. by H. Hardy, *The Crooked Timber of Humanity. Chapters in the History of Ideas*. Edited by H. Hardy. New York: Vintage Books (1992; pp. 1–19).

Brandt, Richard. 1958. "Blameworthiness and Obligation." In A. I. Melden, ed., *Essays in Moral Philosophy*. Seattle: University of Washington Press (pp. 3–39).

Brandt, Richard. 1979. *A Theory of the Good and the Right*. Oxford: Oxford University Press.

Callan, Eamon, and White, John. 2003. "Liberalism and Communitarianism." In N. Blake, P. Smeyers, R. Smith, and P. Standish, eds., *The Blackwell Guide to the Philosophy of Education*. Oxford: Blackwell (pp. 95–109).

Campbell, Richmond, and Sowden, Lanning, eds., 1985. *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*. Vancouver: The University of British Columbia Press.

Chisholm, Roderick. 1986. *Brentano and Intrinsic Value*. Cambridge: Cambridge University Press.
Christman, John. 1987. "Autonomy: A Defense of the Split-Level Self," *The Southern Journal of Philosophy* XXV: 281–93.
Christman, John. 1991. "Autonomy and Personal History," *Canadian Journal of Philosophy* 21: 1–24.
Clarke, Randolph. 1996. "Agent Causation and Event Causation in the Production of Free Action," *Philosophical Topics* 24: 19–48.
Clarke, Randolph. 2003. *Libertarian Accounts of Free Will*. New York: Oxford University Press.
Cocking, Dean, and Kennett, Jeanette. 1998. "Friendship and the Self," *Ethics* 108: 502–27.
Conee, Earl. 1982. "Against Moral Dilemmas." In Christopher Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press (1987; pp. 239–49).
Copp, David. 1997. "The Ring of Gyges: Overridingness and the Unity of Reason." In A. F. Paul, F. D. Miller, Jr., and J. Paul, eds., *Self-Interest*. Cambridge: Cambridge University Press (pp. 86–106).
Cuypers, Stefaan E. 2000*a*. "Autonomy beyond Voluntarism: In Defense of Hierarchy," *Canadian Journal of Philosophy* 30: 225–56.
Cuypers, Stefaan E. 2000*b*. "Alfred Mele's Voluntaristic Conception of Autonomy." In T. van den Beld, ed., *Moral Responsibility and Ontology*. Dordrecht: Kluwer (pp. 259–70).
Cuypers, Stefaan E. 2001. *Self-Identity and Personal Autonomy: An Analytical Anthropology*. Aldershot: Ashgate.
Cuypers, Stefaan E. 2004*a*. "Critical Thinking, Autonomy and Practical Reason," *Journal of Philosophy of Education* 38: 75–90.
Cuypers, Stefaan E. 2004*b*. "The Trouble with Harry: Compatibilist Free Will Internalism and Manipulation," *Journal of Philosophical Research* 29: 235–54.
Cuypers, Stefaan E. 2006. "The Trouble with Externalist Compatibilist Autonomy," *Philosophical Studies* 129: 171–96.
Cuypers, Stefaan E. 2008. "Educating for Authenticity: The Paradox of Moral Education Revisited." In. H. Siegel, ed., *The Oxford Handbook of Philosophy of Education*. New York: Oxford University Press.
Cuypers, Stefaan E., and Bonnett, Michael. 2003. "Autonomy and Authenticity in Education." In N. Blake, P. Smeyers, R. Smith, and P. Standish, eds., *The Blackwell Guide to the Philosophy of Education*. Oxford: Blackwell (pp. 326–40).
Darling, John. 1994. *Child-Centred Education: And Its Critics*. London: Paul Chapman.
Darling, John, and Nordenbo, Sven Erik. 2003. "Progressivism." In N. Blake, P. Smeyers, R. Smith, and P. Standish, eds., *The Blackwell Guide to the Philosophy of Education*. Oxford: Blackwell (pp. 288–308).
Dearden, Robert F. 1972. "Autonony and Education." In R. Dearden, P. Hirst and R. Peters, eds., *Education and the Development of Reason*. London: Routledge and Kegan Paul (pp. 58–75).
Dennett, Daniel. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press.
De Ruyter, Doret J. 2003. "The Importance of Ideals in Education," *Journal of Philosophy of Education* 37: 467–82.
De Ruyter, Doret J. 2006. "Whose Utopia/ Which Ideals? About the Importance of Societal and Personal Ideals in Education." In M. A. Peters and J. Freeman-Moir, eds., *Edutopias: New Utopian Thinking in Education*. Rotterdam: Sense (pp. 163–174).
De Ruyter, Doret J. 2007. "Ideals, Education and Happy Flourishing," *Educational Theory* 57: 23–34.

De Ruyter, Doret J., and Conroy, Jim. 2002. "The Formation of Identity: The Importance of Ideals," *Oxford Review of Education* 28: 509–22.
Donagan, Alan. 1984. "Consistency in Rationalist Moral Systems." In Christopher Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press (1987; pp. 271–90).
Double, Richard. 1989. "Puppeteers, Hypnotists, and Neurosurgeons," *Philosophical Studies* 56: 163–73.
Double, Richard. 1991. *The Non-Reality of Free Will*. Oxford: Oxford University Press.
Double, Richard. 2004. "The Ethical Advantages of Free Will Subjectivism," *Philosophy and Phenomenological Research* 69: 411–22.
Edyvane, Derek. 2003. "Against Unconditional Love," *Journal of Applied Philosophy* 20: 59–75.
Ekstrom, Laura Waddell. 2000. *Free Will: A Philosophical Study*. Boulder: Westview.
Elster, Jon. 1984. *Ulysses and the Sirens. Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.
Emmet, Dorothy. 1994. *The Role of the Unrealisable. A Study in Regulative Ideals*. New York: St. Martin's.
Feinberg, Joel. 1980. "The Child's Right to an Open Future." In W. Aiken and H. LaFollette, eds., *Whose Child? Children's Rights, Parental Authority and State Power*. Totowa, NJ: Rowman & Littlefield (pp. 124–53).
Feinberg, Joel. 1986. *Harm to Self*. New York: Oxford University Press.
Feldman, Fred. 1986. *Doing The Best We Can*. Dordrecht: Reidel Publishing.
Feldman, Fred. 1992. *Confrontations with the Reaper*. New York: Oxford University Press.
Feldman, Fred. 1995. "Adjusting Utility for Justice: A Consequentialist Reply to the Objection from Justice." In Fred Feldman, ed., *Utilitarianism, Hedonism, and Desert*. Cambridge: Cambridge University Press (1997; pp. 154–74).
Feldman, Fred. 2004. *Pleasure and the Good Life. Concerning the Nature, Varieties, and Plausibility of Hedonism*. Oxford: Clarendon.
Fischer, John Martin. 1987. "Responsiveness and Moral Responsibility." In F. Schoeman, ed., *Responsibility, Character, and the Emotions*. Cambridge: Cambridge University Press (pp. 81–106).
Fischer, John Martin. 1994. *The Metaphysics of Free Will*. Oxford: Blackwell.
Fischer, John Martin. 2004. "Responsibility and Manipulation," *The Journal of Ethics* 8: 145–77.
Fischer, John Martin, and Ravizza, Mark. 1994. "Responsibility and History," *Midwest Studies in Philosophy* 19: 430–51.
Fischer, John Martin, and Ravizza, Mark. 1998. *Responsibility and Control: An Essay on Moral Responsibility*. Cambridge: Cambridge University Press.
Foot, Philippa. 1978. "Are Moral Considerations Overriding?" In *Virtues and Vices*. Oxford: Blackwell (pp. 181–88).
Frankfurt, Harry. 1969. "Alternate Possibilities and Moral Responsibility," *The Journal of Philosophy* 66: 829–39. Reprinted in H. Frankfurt. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press (pp. 1–10).
Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of A Person," *The Journal of Philosophy* 68: 5–20. Reprinted in H. Frankfurt. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press (pp. 11–25).
Frankfurt, Harry. 1975. "Three Concepts of Free Action," *Proceedings of the Aristotelian Society,* supp. vol. IL: 113–25. Reprinted in H. Frankfurt. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press (pp. 47–57).

Frankfurt, Harry. 1978. "The Problem of Action," *American Philosophical Quarterly* 15: 157–62. Reprinted in H. Frankfurt. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press (pp. 69–79).

Frankfurt, Harry. 1982. "The Importance of What We Care About," *Synthese* 53: 257–72. Reprinted in Harry Frankfurt. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press (pp. 80–94).

Frankfurt, Harry. 1987. "Identification and Wholeheartedness." In F. D. Schoeman, ed., *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. New York: Cambridge University Press (pp. 27–45). Reprinted in Harry Frankfurt. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press (pp. 159–76).

Frankfurt, Harry. 1992. "On the Usefulness of Final Ends." Reprinted in H. Frankfurt. 1999. *Neccesity, Volition, and Love*. Cambridge: Cambridge University Press (pp. 82–94).

Frankfurt, Harry. 1994. "Autonomy, Necessity, and Love." In Hans Friedrich Fulda and Rolf-Peter Horstmann, eds., *Vernunftbegriffe in der Moderne*. Stuttgart: Klett-Cotta (1994; pp. 433–47). Reprinted in H. Frankfurt. 1999. *Neccesity, Volition, and Love*. Cambridge: Cambridge University Press (pp. 129–41).

Frankfurt, Harry. 1999. "On Caring." In H. Frankfurt, *Neccesity, Volition, and Love*. Cambridge: Cambridge University Press (pp. 155–80).

Frankfurt, Harry. 2002. "Reply to John Martin Fischer." In S. Buss and L. Overton, eds., *Contours of Agency. Essays on Themes from Harry Frankfurt*. Cambridge, MA: MIT Press (pp. 27–31).

Frankfurt, Harry. 2003. "Some Thoughts Concerning PAP." In M. McKenna and D. Widerker, eds., *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities*. Aldershot: Ashgate (pp. 339–45).

Gauthier, David. 1986. *Morals by Agreement*. Oxford: Clarendon.

Ginet, Carl. 1983. "In Defense of Incompatibilism," *Philosophical Studies* 44: 391–400.

Ginet, Carl. 1990. *On Action*. Cambridge: Cambridge University Press.

Ginet, Carl. 2000. "The Epistemic Requirements for Moral Responsibility," *Philosophical Perspectives* 14: 267–77.

Glannon, Walter. 1998. "Moral Responsibility and Personal Identity," *American Philosophical Quarterly* 35: 231–49.

Glannon, Walter. 2002. *The Mental Basis of Responsibility*. Aldershot: Ashgate.

Goetz, Stewart. 1998. "A Noncausal Theory of Agency," *Philosophy and Phenomenological Research* 49: 303–16.

Gowans, Christopher. 1987. "Introduction: The Debate on Moral Dilemmas." In Christopher Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press (pp. 3–33).

Green, O. H. 1997. "Is Love an Emotion?" In Roger E. Lamb, ed., *Love Analyzed*. Boulder: Westview (pp. 209–24).

Gutmann, Amy. 1980. "Children, Paternalism, and Education: A Liberal Argument," *Philosophy and Public Affairs* 9: 338–58.

Haji, Ishtiyaque. 1989. "The Compliance Problem," *Pacific Philosophical Quartely* 70: 105–21.

Haji, Ishtiyaque. 1998. *Moral Appraisability. Puzzles, Proposals, and Perplexities*. New York: Oxford University Press.

Haji, Ishtiyaque. 2001*a*."Control Conundrums: Modest Libertarianism, Responsibility, and Explanation," *Pacific Philosophical Quarterly* 82: 178–200.

Haji, Ishtiyaque. 2001*b*. "Self-Deception and Blameworthiness," *Journal for the Theory of Social Behavior* 31: 279–95.

Haji, Ishtiyaque. 2002. *Deontic Morality and Control*. Cambridge: Cambridge University Press.

Haji, Ishtiyaque. 2003*a*. "The Emotional Deparvity of Psychopaths and Culpability," *Legal Theory* 9: 63–82.
Haji, Ishtiyaque. 2003*b*. "Determinism and its Threat to the Moral Sentiments," *The Monist* 86: 244–62.
Haji, Ishtiyaque. 2004. "Freedom, Hedonism, and the Intrinsic Value of Lives," *Philosophical Topics* 32: 131–51.
Haji, Ishtiyaque. forthcoming (a). "Incompatibilism's Threat to Worldly Value: Source Incompatibilism, Desert, and Pleasure," *Philosophy and Phenomenological Research*.
Haji, Ishtiyaque. forthcoming (b). *Freedom and Value. Freedom's Influence on Welfare and Worldly Value. Dordrecht: Springer.*
Haji, Ishtiyaque, and Cuypers, Stefaan E. 2001. "Libertarian Free Will and CNC Manipulation," *Dialectica* 55: 221–38.
Harman, Gilbert. 1967. "Toward a Theory of Intrinsic Value." *Journal of Philosophy* 64: 792–804.
Henson, Richard. 1979. "What Kant Might Have Said: Moral Worth and the Overdetermination of Dutiful Action," *Philosophical Review* 88: 39–54.
Heyting, Frieda. 2004. "Beware of Ideals in Education," *Journal of Philosophy of Education* 38: 241–47.
Hobbes, Thomas. 1650. *The Elements of Law, Natural and Politic*. Edited by and with an introduction by J. C. A. Gaskin. Oxford: Oxford University Press (1994).
Hobbes, Thomas. 1651. *Leviathan*. Edited with an introduction by C. B. Macpherson. Harmondsworth: Penguin Books (1985).
Hoffman, Martin. 1976. "Empathy, Role Taking, Guilt and Development of Altruistic Motives." In T. Likona, ed., *Moral Development and Behavior: Theory Research and Social Issues*. New York: Holt, Rinehart, Winston (pp. 124–43).
Honderich, Ted. 1993. *How Free Are You?* Oxford: Oxford University Press.
Hudson, Stephen D. 1986. *Human Character and Morality*. London: Routledge and Kegan Paul.
Hume, David. 1739. *A Treatise of Human Nature*. Edited by Lewis Selby-Bigge, Oxford: Clarendon Press (1975).
Hume, David. 1748. *Enquiry Concerning Human Understanding*. Edited by E. Steinberg, Indianapolis: Hackett (1981).
Huxley, Aldous. 1937. *Ends and Means. An Enquiry into the Nature of Ideals and into the Methods Employed for their Realization*. London: Chatto and Windus.
Kagan, Shelly. 1998. *Normative Ethics*. Boulder: Westview.
Kane, Robert. 1996. *The Significance of Free Will*. New York: Oxford University Press.
Kane, Robert. 2000*a*. "The Dual Regress of Free Will and the Role of Alternative Possibilities," *Philosophical Perspectives* 14: 57–79.
Kane, Robert. 2000*b*. "Non-Constraining Control and the Threat of Social Conditioning," *The Journal of Ethics* 4: 401–03.
Kant, Immanuel. 1795. *Groundwork of the Metaphysic of Morals*. Translated and analysed by H. J. Paton, New York: Harper and Row (1964).
Kapitan, Tomis. 2000. "Autonomy and Manipulated Freedom," *Philosophical Perspectives* 14: 81–103.
Kolodny, Niko. 2003. "Love as Valuing a Relationship," *The Philosophical Review* 112: 135–89.
Kraut, Richard. 1986. "Love De Re." *Midwest Studies in Philosophy* 10: 413–30.
Lamb, James. 1993. "Evaluative Compatibilism and the Principle of Alternate Possibilities," *The Journal of Philosophy* 90: 517–27.
Lamb, Roger E. 1997. "Love and Rationality." In Roger E. Lamb, ed., *Love Analyzed*. Boulder: Westview (pp. 23–47).

Lemmon, E. J. 1962. "Moral Dilemmas." In Christopher Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press (1987; pp. 101–14).
Lemos, Noah. 1994. *Intrinsic Value*. Cambridge: Cambridge University Press.
Locke, Don. 1975. "Three Concepts of Free Action I," *Proceedings of the Aristotelian Society*, Sup. Vol. 49: 95–112.
Mackie, J. L. 1977. *Ethics. Inventing Right and Wrong*. Harmondsworth: Penguin.
Marcus, Ruth Barkan. 1980. "Moral Dilemmas and Consistency." In Christopher Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press (1987; pp. 188–204).
Marples, Roger, ed. 1999. *The Aims of Education*. London: Routledge.
McConnell, Terrance. 1978. "Moral Dilemmas and Consistency in Ethics." In Christopher Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press (1987; pp. 154–73).
McKenna, Michael. 2004*a*. "Responsibility and Globally Manipulated Agents," *Philosophical Topics* 32: 169–92.
McKenna, Michael. 2004*b*. "Compatibilism." In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy (Summer Edition)*, URL = http://plato.stanford.edu/archives/sum2004/entries/compatibilism/.
McKenna, Michael. 2005*a*. "Where Frankfurt and Strawson Meet," *Midwest Studies in Philosophy* 29: 163–80.
McKenna, Michael. 2005*b*. "The Relationship Between Autonomous and Morally Responsible Agency." In James Taylor, ed., *Personal Autonomy: New Essays On Personal Autonomy and its Role in Contemporary Moral Philosophy*. Cambridge: Cambridge University Press (pp. 205–34).
McKenna, M. forthcoming. "A Hard-line Reply to Pereboom's Four-Case Manipulation Argument," *Philosophy and Phenomenological Research*.
Mele, Alfred. 1987. *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. New York: Oxford University Press.
Mele, Alfred. 1992. *Springs of Action: Understanding Intentional Behavior*. New York: Oxford University Press.
Mele, Alfred. 1993. "History and Personal Autonomy," *Canadian Journal of Philosophy* 23: 271–80.
Mele, Alfred. 1995. *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press.
Mele, Alfred. 2000. "Goal-Directed Action: Teleological Explanations, Causal Theories, and Deviance," *Philosophical Perspectives* 14: 279–300.
Mele, Alfred. 2003. *Motivation and Agency*. New York: Oxford University Press.
Mele, Alfred. 2005. "A critique of Pereboom's 'Four-Case Argument' for Incompatibilism," *Analysis* 65: 75–80.
Mele, Alfred. 2006. *Free Will and Luck*. New York: Oxford University Press.
Mill, John Stuart. 1863. *Utilitarianism*. New York: Macmillan, 1989.
Milo, Ronald D. 1984. *Immorality*. Princeton: Princeton University Press.
Moore, George E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.
Mulhall, Stephen, and Swift, Adam. 1996. *Liberals and Communitarians* (2nd ed.). Oxford: Blackwell.
Murphy, Jeffrie, and Hampton, Jean. 1988. *Forgiveness and Mercy*. Cambridge: Cambridge University Press.
Murris, Karin. 2000. "Can Children Do Philosophy?," *Journal of Philosophy of Education* 34: 261–79.
Noddings, Nel, and Slote, Michael. 2003. "Changing Notions of the Moral and of Moral Education." In N. Blake, P. Smeyers, R. Smith, and P. Standish, eds., *The Blackwell Guide to the Philosophy of Education*. Oxford: Blackwell (pp. 341–55).

Oakley, Justin. 1992. *Morality and the Emotions*. New York: Routledge.
O'Connor, Timothy. 2000. *Persons and Causes*. New York: Oxford University Press.
O'Connor, Timothy. 2002. "The Agent as Cause." In Robert Kane, ed., *Free Will*. Oxford: Blackwell (pp. 196–205).
Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon.
Paton, H. J. 1964. *Immanuel Kant's Groundwork of the Metaphysic of Morals* (1795). Translated and analysed. New York: Harper and Row.
Paul, Ellen Frankel, Miller Jr., Fred D., Paul, Jeffrey, and Ahrens, John, eds, 1988. *The New Social Contract: Essays on Gauthier*. Oxford: Basil Blackwell.
Pereboom, Derk. 1995. "Determinism *al Dente*," *Noûs* 29: 21–45.
Pereboom, Derk. 2000. "Alternative Possibilities and Causal Histories," *Philosophical Perspectives* 24: 119–37.
Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.
Pereboom, Derk. 2002. "Living Without Free Will: The Case for Hard Incompatibilism." In Robert Kane, ed., *The Oxford Handbook of Free Will*. New York: Oxford University Press (pp. 477–88).
Peters, Richard S. 1963. "Reason and Habit: the Paradox of Moral Education." In Richard S. Peters, ed., *Psychology and Ethical Development. A Collection of Articles on Psychological Theories, Ethical Development and Human Understanding*. London: George Allen & Unwin (1974; pp. 265–80).
Peters, Richard S. 1973. "Freedom and the Development of the Free Man." In Richard S. Peters, ed., *Psychology and Ethical Development. A Collection of Articles on Psychological Theories, Ethical Development and Human Understanding*. London: George Allen & Unwin (1974; pp. 336–59).
Pettit, Philip. 1997. "Love and its Place in Moral Discourse." In Roger E. Lamb, ed., *Love Analyzed*. Boulder: Westview (pp. 153–63).
Purdy, Laura M. 1992. *In their Best Interest? The Case Against Equal Rights for Children*. Ithaca, NY: Cornell University Press.
Rescher, Nicholas. 1966. *Distributive Justice*. New York: Bobbs-Merrill.
Rescher, Nicholas. 1987. *Ethical Idealism. An Inquiry into the Nature and Function of Ideals*. Berkeley: University of California Press.
Ross, W. D. 1930. *The Right and the Good*. Oxford: Oxford University Press.
Rousseau, Jean-Jacques. 1762. *Emile or On Education*. Introduction, Translation, and Notes by A. Bloom, New York: BasicBooks (1979).
Scheffler, Israel. 1971. "Moral Education and the Democratic Ideal." In I. Scheffler, ed., *Reason and Teaching*. London: Routledge and Kegan Paul (1973; pp. 136–45).
Shatz, David. 1986. "Free Will and the Structure of Motivation," *Midwest Studies in Philosophy* 10: 451–82.
Siegel, Harvey. 1987. *Relativism Refuted. A Critique of Contemporary Epistemological Relativism*. Dordrecht: Reidel.
Siegel, Harvey. 1988. *Educating Reason. Rationality, Critical Thinking, and Education*. New York: Routledge.
Siegel, Harvey. 1991. "Indoctrination and Education." In B. Spiecker and R. Straughan, eds., *Freedom and Indoctrination in Education. International Perspectives*. London: Cassell (pp. 30–41).
Siegel, Harvey. 1997. *Rationality Redeemed? Further Dialogues on an Educational Ideal*. New York: Routledge.
Siegel, Harvey. 1998. "Knowledge, Truth and Education." In D. Carr, ed., *Education, Knowledge and Truth. Beyond the Post-Modern Impasse*. London: Routledge (pp. 19–36).
Siegel, Harvey. 2003. "Cultivating Reason." In R. Curren, ed., *A Companion to the Philosophy of Education*. Oxford: Blackwell (pp. 305–19).

Siegel, Harvey. 2005. "Neither Humean nor (Fully) Kantian Be: Reply to Cuypers," *Journal of Philosophy of Education* 39: 535–47.
Sinnott-Armstrong, Walter. 1988. *Moral Dilemmas*. Oxford: Basil Blackwell.
Skinner, B.F. 1948. *Walden Two*. Englewood Cliffs: Prentice-Hall (1976).
Slote, Michael. 1983. "Admirable Immorality." In M. Slote, ed., *Goods and Virtues*. Oxford: Clarendon (pp. 77–107).
Slote, Michael. 1992. *From Morality to Virtue*. New York: Oxford University Press.
Smart, J. J. C. 1973. "An Outline of a System of Utilitarian Ethics." In J. J. C. Smart and Bernard Williams, ed., *Utilitarianism: For and Against*. Cambridge: Cambridge University Press (pp. 1–74).
Smeyers, Paul, and Wringe, Colin. 2003. "Adults and Children." In N. Blake, P. Smeyers, R. Smith, and P. Standish, eds., *The Blackwell Guide to the Philosophy of Education*. Oxford: Blackwell (pp. 311–25).
Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell.
Smith, Michael. 1996. "The Argument for Internalism: Reply to Miller," *Analysis* 56: 175–84.
Snook, I. A., ed., 1972. *Concepts of Indoctrination. Philosophical Essays*. London: Routledge & Kegan Paul.
Spiecker, Ben and Straughan, Roger, eds., 1991. *Freedom and Indoctrination in Education. International Perspectives*. London: Cassell.
Standish, Paul. 1999. "Education Without Aims?" In R. Marples, ed., *The Aims of Education*. London: Routledge (pp. 35–49).
Standish, Paul. 2003. "The Nature and Purposes of Education." In R. Curren, ed., *A Companion to the Philosophy of Education*. Oxford: Blackwell (pp. 221–31).
Stocker, Michael. 1990. *Plural and Conflicting Values*. Oxford: Clarendon.
Strawson, Galen. 1986. *Freedom and Belief*. Oxford: Clarendon.
Strawson, Peter F. 1962. "Freedom and Resentment," *Proceedings of the British Academy* 48: 1–25. Reprinted in Gary Watson, ed., 1982. *Free Will*. Oxford: Oxford University Press (pp. 59–80).
Sumner, L. W. 1996. *Welfare, Happiness, and Ethics*. Oxford: Clarendon.
Thalberg, Irving. 1978. "Hierarchical Analyses of Unfree Action," *Canadian Journal of Philosophy* 8: 211–26.
Thomas, Laurence. 1987. "Friendship," *Synthese* 72: 217–36.
Twain, Mark. 1884. *The Adventures of Huckleberry Finn*. London, England: Penguin Classics (1985).
Vallentyne, Peter. 1991. Ed. *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement*. Cambridge: Cambridge University Press.
van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.
Van Fraassen, Bas. 1973. "Values and the Heart's Command." In Christopher Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press (1987; pp. 138–53).
Velleman, J. David. 1992. "What Happens When Someone Acts?" *Mind* 101: 461–81.
Velleman, J. David. 1999. "Love as a Moral Emotion," *Ethics* 109: 338–74.
Velleman, J. David. n.d. "Frankfurt on Love and Duty."
Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
Waller, Bruce. 1990. *Freedom Without Responsibility*. Philadelphia: Temple University Press.
Warfield, Ted. 1996. "Determinism and Moral Responsibility are Incompatible," *Philosophical Topics* 24: 215–26.

Watson, Gary. 1975. "Free Agency," *Journal of Philosophy* 72: 205–20. Reprinted in Gary Watson, ed., 1982. *Free Will*. Oxford: Oxford University Press (pp. 96–110).
Watson, Gary. 1987*b*. "Free Action and Free Will," *Mind* 96: 145–72.
Watson, Gary. 1999. "Soft Libertarianism and Hard Compatibilism," *The Journal of Ethics* 3: 351–65.
White, John. 1982. *The Aims of Education Restated*. London: Routledge and Kegan Paul.
White, John. 1990. *Education and the Good Life. Beyond the National Curriculum*. London: Kogan Page.
White, John. 1999. "In Defence of Liberal Aims in Education." In R. Marples, ed., *The Aims of Education*. London: Routledge (pp. 185–200).
White, John. 2003. "Five Critical Stances Towards Liberal Philosophy of Education in Britain," (with responses by W. Carr, R. Smith, P. Standish, and T. H. McLaughlin) *Journal of Philosophy of Education* 37: 147–84.
Wiggens, David. 1973. "Towards a Reasonable Libertarianism." In T. Honderich, ed., *Essays on Freedom of Action*. London: Routledge and Kegan Paul (pp. 33–61).
Williams, Bernard. 1973. "Ethical Consistency." In Christopher Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press (1987; pp. 115–37).
Williams, Bernard. 1976*a*. "Moral Luck," *Proceedings of the Aristotelian Society*, Sup. Vol. 50: 115–35.
Williams, Bernard. 1976*b*. "Persons, Character and Morality." In A. Oksenberg Rorty, ed., *The Identities of Persons*. Berkeley: University of California Press (pp. 197–215).
Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. London: Fontana.
Winch, Christopher. 1999. "Autonomy as an Educational Aim." In R. Marples, ed., *The Aims of Education*. London: Routledge (pp. 74–84).
Winch, Christopher. 2002. "Strong Autonomy and Education," *Educational Theory* 52: 27–41.
Winch, Christopher, and Gingell, John. 1999. *Key Concepts in the philosophy of Education*. London: Routledge.
Wolf, Susan. 1982. "Moral Saints," *The Journal of Philosophy* 19: 419–39.
Wolf, Susan. 1990. *Freedom Within Reason*. New York: Oxford University Press.
Wolff, Robert Paul. 1973. *The Autonomy of Reason*. New York: Harper and Row.
Yaffe, Gideon. 2003. "Indoctrination, Coercion and Freedom of Will," *Philosophy and Phenomenological Research* LXVII: 335–56.
Zimmerman, David. 1999. "Born Yesterday: Personal Autonomy for Agents Without a Past," *Midwest Sudies in Philosophy:* 236–66.
Zimmerman, David. 2002. "Reasons-Responsiveness and Ownership-of-Agency: Fischer and Ravizza's Historicist Theory of Responsibility," *The Journal of Ethics* 6: 199–234.
Zimmerman, David. 2003*a*. "That Was Then, This Is Now: Personal History vs. Psychological Structure in Compatibilist Theories of Autonomous Agency," *Noûs* 37: 638–71.
Zimmerman, David. 2003*b*. "Why Richard Brandt Does Not Need Cognitive Psychotherapy, and Other Glad News about Idealized Preference Theories in Meta-Ethics," *The Journal of Value Inquiry* 37: 373–94.
Zimmerman, Michael J. 1988. *An Essay on Moral Responsibility*. Totowa, NJ: Rowman & Littlefield.
Zimmerman, Michael J. 1996. *The Concept of Moral Obligation*. Cambridge: Cambridge University Press.
Zimmerman, Michael J. 1997. "A Plea for Accuses," *American Philosophical Quarterly* 34: 229–43.

Zimmerman, Michael J. 2001. *The Nature of Intrinsic Value*. Lanham, MD: Rowman & Littlefield.
Zimmerman, Michael J. 2002. "Taking Luck Seriously," *Journal of Philosophy* 99: 553–76.
Zimmerman, Michael J. 2004. "Another Plea for Excuses," *American Philosophical Quarterly* 41: 259–66.
Zimmerman, Michael J. 2007. "Feldman on the Nature and Value of Pleasure," *Philosophical Studies* 136: 425–37.

Index

A

action, 27
  causal theory of, 27
  wholly intrinsically motivated, 131, 133
  appraisability for, 134–44
actional springs, unsheddable, 12
Adams, Robert, 176
agent, 2
  amoral, 134
  magical, 42, 46–47, 60
  moral, 2, 22, 24, 33, 70–71
  morally normatively responsible, 3, 4, 20–21, 23, 29, 35, 43, 49, 70–71, 88
  normatively responsible, 6, 108–109, 111
agent-causation, 11, 17–18, 91
alternatives, genuine, 18
Archard, David, 65, 66, 69
aretaic appraisals, 128, 142–44
Aristotle, 8, 131, 143, 169, 186
Arpaly, Nomy, 129–131
attitudinal hedonism, 175
  object-worthy, 180–81
  simple, 176
  subject's desert-adjusted, 179–80
  truth-adjusted, 176–77
authenticity, 3, 5, 6, 9, 15, 19, 24–26, 42, 86–87
  educational, 2, 4, 14, 63–74
  plain, 22, 23, 33, 70, 74, 87
  relational, 3–5, 22, 23, 28, 35, 41, 55, 62, 69, 70–72, 87, 88

B

behavior, lovable, 115, 120–22
  bearing of hard incompatibilism on, 157–58
  and commendability, 121–26
blameworthiness, 7, 57
  moral, 5
Benn, Stanley, 64, 78
Brandt, Richard, 202, 203

C

causal routes, 4
  deviant, 4, 14
  normal, 16, 28, 31
censurability, 5, 7, 108
childhood, 64–65
  free-standing conception, 65
  stage conception, 65
Chisholm, Roderick, 178
Christman, John, 82
Clarke, Randolph, 11
Cocking, Dean, 117, 123–24
commendability, 5, 7, 107–108, 110, 122, 126
compatibilism, 2, 9, 10, 14, 16–18, 20, 23, 38, 41, 47, 62, 68
control, 8, 24, 30, 58
  CNC, 9, 16, 22, 33, 37
  directional, 9
  dual directional, 18
  guidance, 189–90
  hierarchical, 9, 13, 17, 47, 72
  ultimate, 16–17
critical thinker, 3, 79, 83, 88
  autonomous, 3, 75, 83, 86–88
  proto, 79, 81, 82
critical thinking, 75
  and well-being, 186

D

Dearden, Robert, 64, 77
De Ruyter, Doret, 162–63
desert 177, 178

axiological principles involving, 178–79
bases of, 177–78
desire, 13
first-order, 13, 38–40
intrinsic, 131
irresistible, 24, 36
second-order, 13, 38–40
wholly intrinsic, 131
determinism, 2, 7, 9, 10, 38–39
Dewey, John 167
duty, acting from, 4, 113–14, 142–44, 154

E

education, 2, 63
aims of, 87, 162–163
ideals of, 64, 76–77, 79, 162–163
liberal paradigm, 170–71
non-liberal paradigm, 170–71
overarching aims of, 164–65
skepticism concerning aims of, 165–69
educational authenticity, objection, 2, 14, 64, 67–68
reply to, 71–72
Edyvane, Derek, 151–52
Ekstrom, Laura, 10
evaluative scheme, 3, 20, 25, 42, 50, 54, 58
acceptable modifications to, 29
authentic, 21, 26
evolved, 21, 29–31, 49
initial, 4, 21, 26–27, 29, 34, 52, 55, 58
externalism, 32, 42, 45, 53–55, 57

F

Feinberg, Joel, 23, 24, 26, 33, 69
Feldman, Fred, 175, 176, 177, 179, 182
Fischer, John, 9, 71, 89, 164
forgiveness, 93, 96
four-case argument, 38–40
response to, Fischer and Ravizza's, 190–93
response to, McKenna's, 212–17
response to, ours, 40–41
Frankfurt, Harry, 9
and blameworthiness, 130
and hierarchical identification, 13
and internalism, 46, 218
and love's demands, 116–117
and love's objects, 153
and the necessities of love, 58
and non-overridingness of moral obligation, 112–13
and the *Participation Principle,* 43–44
Frankfurt-type example, 193–94

G

Gauthier, David, 84, 85
gratitude, 97–98, 102
guilt, 98–99, 101, 102–106
Gutman, Amy, 69

H

hard incompatibilism, 5, 90–91
Hoffman, Martin, 204

I

identification, hierarchical, 13, 44, 51–52
incompatibilism, 5, 6, 9–10, 16–17, 19, 23, 62, 68
indoctrination, 2, 4, 6, 15, 22–24, 32, 63, 65, 67
objection, 3, 14, 75, 80
problem of, 74–88
response to, Seigel's, 80–81
response to, ours, 86–88
internalism, 42, 45, 49, 51, 53–55, 201
intrinsic value states, basic, 174–75

K

Kane, Robert, 9, 33
and CNC control, 9–10
and ultimate control, 16–17, 89
Kapitan, Tomis, 35, 36
Kennett, Jeanette, 117, 123–24
Kolodny, Niko, 117, 119
Kraut, Richard, 119

L

Lamb, Roger, 121
Lemos, Noah, 178
libertarianism, 2, 5, 10–11, 16, 18
love, 4–7
acting from, 5, 108, 147–48, 154–55
analyses of, 116
bearing of determinism on, 158–62
bearing of hard incompatibilism on, 118–21
connection with well-being, 183–85
partiality of, 151–53
requirements of, 116, 121, 122
value of, 116–17
Velleman's account of, 149–51

M

magical agent, objection, 47–49

reply to, 50
manipulation, 1, 6, 14–15, 19, 33–35, 38, 42, 56, 62
CNC, 9–10, 18, 40
global, 11–12, 21, 36, 43–49, 51, 59
objection, 1–2, 9–10
Marples, Roger, 70
Mele, Alfred, 12, 13, 201, 230n.11
against Arpaly, 58–59
and control over emotions, 99–100
and deliberative control, 29–30
and externalism, 45
and the four-case argument, 219
and the manipulation problem, 206–210
and wholly intrinsically motivated action, 131–32
McKenna, Michael, 6, 33, 41
and the four-case argument, 212–17
against externalism, 47–48, 54
Mill, John Stuart, 113
Moore, G. E., 177, 178

N

necessity, volitional, 58
Noddings, Nel, 170

O

O'Connor, Timothy, 11, 19
Obligation, moral, 94
incompatibility with determinism, 94–95
overridingness, 172–73

P

Participation Principle, 43, 46, 49–51, 53–54, 57, 60–61
Parfit, Derek, 174
paternalism, 2, 4, 22–23, 65–66, 68–69, 71–72
Pereboom, Derk, 5, 106
and analogs of attitudes, 96–97, 100–102
and forgiveness, 93
and the four-case argument, 38–40
and guilt, 104, 105
and hard incompatibilism, 90–92
and love, 117–19, 156–57
and moral reform and education, 95
Peters, Richard, 64, 77, 78
Pettit, Philip, 145–47
praiseworthiness, 7, 57
moral, 5, 113
prisoner's dilemma, 84–85
progressivism, 64
pro-attitude, inauthentic, 12, 25, 70
Purdy, Laura, 66

R

rationality, 77, 84–85
Ravizza, Mark, 9, 71, 89, 164
reasons-responsiveness, 9, 47
moderate, 38–40, 189–90
remorse, 98, 103, 106
responsibility, 1, 4, 7, 12, 22, 38, 42, 54, 57, 67, 75, 86
agency condition of, 8, 71
analysis of, 8, 19–20
authenticity condition of, 8
belief requirement of, 128–140, 145
connection with well-being, 185–86
freedom condition of, 8
ledger view of, 135
normative, 108–111
taking, Fischer and Ravizza's account, 190, 225
Ross, W. D., 177
Rousseau, J. J., 65

S

Siegel, Harvey, 2, 14, 74–75, 83, 85–86, 88
and the reasons conception of critical thinking, 76–79
reply to the indoctrination objection, 80–81
Skinner, B. F., 1
Smart, J. J. C., 178
sorrow, 98, 101, 102, 103, 106
Slote, Michael, 111–12, 170
Standish, Paul, 165–69
Strawson, Peter, 71, 96, 134
Sumner, L. W., 176

T

Thomas, Laurence, 117

V

Velleman, David, 44, 124–25, 149–51
value, intrinsic, 6
virtue, acting from, 142–44

W

Wallace, Jay, 9
White, John, 170–71
Williams, Bernard, 108, 123
Wolf, Susan, 200

worlds, intrinsic value of, 180–83

Y
Yaffee, Gideon, 195–97, 200

Z
Zimmerman, David, 45, 63, 201–205
Zimmerman, Michael, 125, 131, 135, 181, 198