THESIS

LIBRARY
Michigan State
University

This is to certify that the

dissertation entitled

COMPATIBILISM WITHOUT UTILITARIANISM:
MORAL RESPONSIBILITY IN A DETERMINISTIC WORLD

presented by

NICHOLAS JOHN DIXON

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Philosophy

Bruce L. Miller

Major professor

Date 12 November 1985

MSU is an Affirmative Action/Equal Opportunity Institution

O-12771



RETURNING MATERIALS:
Place in book drop to remove this checkout from your record. FINES will be charged if book is returned after the date stamped below.

COMPATIBILISM WITHOUT UTILITARIANISM: Moral Responsibility in a Deterministic World

By

Nicholas John Dixon

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Philosophy

1985

3846684

ABSTRACT

COMPATIBILISM WITHOUT UTILITARIANISM: Moral Responsibility in a Deterministic World

By

Nicholas John Dixon

This dissertation offers a compatibilist solution to the apparent conflict - illustrated by the "slippery slope" argument from occasional to universal exculpation - between sufficient causal explanations of human actions, and moral responsibility. In reply to traditional compatibilist solutions, recent incompatibilists have insisted that, if even our desires are caused, we cannot do otherwise than we actually do, and are therefore not responsible. (Chapter I)

Dennett's compatibilist response is a "naturalistic" account of the conditions necessary for responsibility - e.g., rationality and control - which assumes the truth of determinism. His boldest move is one of his replies to the recent incompatibilist argument: even if determinism is true, we can do otherwise in the sense that more than one action is always epistemically possible. (Chapter II)

Chapter III proposes a criterion of responsibility for particular actions, based on our capacity for qualitative evaluations. The incompatibilist wrongly

pictures us as helplessly following our desires, whereas we can in fact determine ourselves by controlling these desires. However, the hard determinist would insist that even this ability to control ourselves is ultimately due to genetic and environmental factors beyond our control. These factors are a matter of moral luck, and Dennett's notion of epistemic possibility does not address this problem.

Nonetheless, Dennett's overall strategy concerning moral luck is successful. To attribute all actions to luck is to misuse the concept, and ignores the impact of our skill on our actions. However, Dennett's adoption of a utilitarian view of punishment and responsibility makes him vulnerable to the slippery slope argument. My criterion, in contrast, shows exactly when it is fair to excuse victims of bad moral luck. The influence of moral luck should not, however, be entirely dropped from our moral framework. (Chapter IV)

Chapter V presents a retributive critique of a purely utilitarian view of punishment and responsibility. Most compatibilists have adopted a utilitarian view because of their inability to find a principled solution to the slippery slope argument. However, compatibilism is in fact quite consistent with a retributive view. This is shown by Strawson's discussion of reactive attitudes, and by my nonutilitarian solution to the slippery slope argument.

TABLE OF CONTENTS

Chapter		Page
I.	The Determinism/Moral Responsibility Debate .	1
	1. The "Slippery Slope"	1
	2. Determinism and Indeterminism	13
	3. Traditional Compatibilist Solutions . . .	20
	a. Causation does not entail compulsion .	21
	b. Responsibility depends on the unimpeded following of our desires	23
	c. Responsibility and punishment	26
	d. "Could have done otherwise"	30
	4. New Incompatibilist Challenges	32
	a. The causation of our desires	33
	(i) The libertarian objection	33
	(ii) The hard determinist objection .	38
	b. Incompatibility at the linguistic level	43
	c. Determinism and alternate possibilities	52
II.	Dennett's Reply to the New Incompatibilists .	60
	A. BRAINSTORMS	61
	1. Levels of Description	61
	2. Diagnosis of Belief in Incompatibility	67
	3. A Tentative Account of Responsibility	70
	B. ELBOW ROOM.	76
	1. Overview	76
	2. The Bugbears	80
	3. The Naturalistic Account	85
	a. Rationality	85
	b. Control	89
	c. Self-determination	94
	d. Effective deliberation	105
	4. "Could Have Done Otherwise"	113
	5. The Justification of our Concept of Responsibility	121
	6. Reflections	126

Chapter		Page
III.	The Necessary and Sufficient Conditions for Responsibility	129
	1. The Slippery Slope Revisited	129
	2. Desires and Compulsion	132
	3. An Analysis of Self-Determination, and a Criterion of Responsibility	137
	4. Applications of the Analysis to Traditional Excuses	154
	a. Ignorance	155
	b. Compulsion	156
	c. Insanity	158
	5. Hard Determinism Revisited	170
	6. Summary	185
IV.	Moral Luck	188
	1. The Threat Posed by Moral Luck	189
	2. Dennett's Reply	201
	a. Constitutive luck	206
	b. Circumstantial luck	208
	c. Consequential luck	210
	3. The Slippery Slope Revisited, Again	214
	4. Extreme Cases of Bad Consequential Luck	219
	5. A Partial Solution to the "Drunken Drivers" Case	228
	6. Morality Without Luck	231
	a. Bad luck and personal causal connections	232
	b. Responsibility without causation	238
	7. Overview	246
V.	Compatibilism and Retribution	250
	1. Utilitarian and Retributive Justifications of Punishment	250
	2. A Critique of a Utilitarian View of Punishment	252
	3. Why Most Compatibilists are Utilitarians	268
	4. Why Compatibilists Need Not be Utilitarians	273

CHAPTER I

THE DETERMINISM/MORAL RESPONSIBILITY DEBATE

1. The "Slippery Slope"

A schizophrenic in a mental hospital is hallucinating uncontrollably. A dose of lithium, or a similar drug, is administered, and within minutes the patient is back to normal, able to converse calmly, and amazed when told how she was behaving so recently. An epileptic flails his arms and legs around wildly during a seizure. We understand his behavior as uncontrollable reflex reactions due to a temporary brain disorder. When such stark physiological dysfunctions are cited as the cause of deviant behavior, we don't hesitate to excuse from responsibility. They "couldn't help" what they were doing.

Our attitude is similar toward those whose deviant actions have clear-cut environmental causes. It has been found that nearly all parents who sexually abuse their children were themselves sexually abused by their parents. Consider also a man in his mid-20's who is convicted of a series of assaults and violent robberies. After an investigation of his past, it is discovered that he was repeatedly beaten by his father, who was eventually jailed

for assaulting his wife. The criminal and his nine siblings were raised by their mother, who encouraged them to shoplift in order to get enough food. In his early teens, he discovered that the easiest way to get money was to join the violent gangs to which many of his neighborhood friends belonged. He became more and more deeply involved in these gangs and their robberies until his conviction. The existence of such environmental causes is unlikely to exempt a malefactor from blame.[1] However, in proportion to the severity of the environmental handicap suffered by the criminal, we are likely to diminish our blame. It seems perfectly appropriate for defense attorneys to bring such background information into light for the purposes of sentencing, if not for conviction itself. While such pleas for diminished responsibility may doubtless be abused, there seem to be clear cases in which they are appropriate.

---

[1] It may be questioned whether these are genuine cases of causation. It is often pointed out, for example, that some victims of sexual abuse do not go on to abuse their children; and that some children with terribly violent home backgrounds "rise above" their origins and become respected professionals. Clearly such cases exist, and so having such a bad environment is at best a necessary condition for becoming a child molester or violent criminal. However, we do sometimes offer a necessary condition as a causal explanation, even if it only creates a low probability of the occurrence of the event it is cited to explain. For example, only 28 percent of those with untreated latent syphilis develop paresis or similar disorders, yet this does not prevent us from citing syphilis in our causal explanation of how they contracted paresis. This example is discussed by Wesley Salmon, in "Determinism and Indeterminism in Modern Science," in Joel Feinberg (ed.), Reason and Responsibility (4th edition, Dickinson, 1978), p. 340.

A further kind of exculpating causation emerges in, for example, some murder trials. In the previous group of cases, the environmental influences were in the distant past, working on the criminal without her ever being aware of them. However, we sometimes consider the immediate "trigger" to action, of which the criminal is perfectly well aware, as a reason for leniency. Thus in most legal systems "crimes of passion" are distinguished from murder motivated by financial gain or other material goals. First degree murder is usually reserved for cold-blooded killing, and is less likely to be charged if the crime was committed in the heat of a rage. Murders committed out of anger, jealousy, frustration, rejection, humiliation, fear, feelings of inferiority and many other emotions are terrible crimes, and never to be condoned. However, it does seem appropriate that we regard them more leniently than callously planned murders done for profit only. A good example is the case of the Michigan woman, recently dramatized in the TV movie "The Burning Bed," who killed her husband after he had repeatedly beaten her for years. It is worth noting that the verdict of "not guilty by reason of insanity" is itself an example of legal and moral insanity. Given the woman's situation - her pleas to her husband, her courtroom battles to become separated, and his persistence in ignoring court orders, returning to live with her, and resuming his assaults on her - her decision

to kill him was not at all irrational (though of course hardly commendable). It is precisely because her action was, to her mind, the only reasonable solution to a terrible situation, that we understand and tend to forgive her. Our leniency is based on the belief that her husband's cruel provocation caused her to kill him. So exculpating causation can operate via the malefactor's rationality, when she is reacting emotionally to events and situations of which she is perfectly well aware.[1]

Excuses based on the causation of our behavior are also common in everyday life. A good friend is unusually abrupt in greeting you one day, and treats you like a stranger. Your offense and resentment at her rudeness will be tempered if she later tells you that she was suffering from awful menstrual pains, or a splitting headache. And not only physical conditions provide excuses. You will be equally forgiving if you later discover that she had just been fired from her job, or learned that her mother had been diagnosed as having irreversible cancer. Causes of behavior which stretch back into the past are also recognized in everyday life as excusing factors. During a classroom discussion of abortion, a student becomes uncharacteristically vehement and abusive in arguing

---

[1]Again, there is always the possibility that such excuses may be abused. In a recent case "pre-menstrual syndrome" (PMS) was admitted as a defense against murder charges. The attempt at such dubious excuses should not, however, blind us to clear cases where the excuse is valid.

against anti-abortionists. Your condemnation of her rudeness will diminish when you learn that as an 11-year-old, she was raped and had an abortion.

What is proven by all the examples so far discussed? It might well be insisted that these are all extreme cases in which causation does indeed diminish or negate responsibility, but that this is precisely because the cases are rare. Most of us, most of the time, are morally responsible for our actions. Most of us have indeed escaped the misfortunes described above: we were not sexually abused by our parents, we were not exposed to violent gangs in our youth, we suffer from no major chemical imbalance in our brains, we have never been exposed to provocation comparable to domestic violence, etc. However, we <u>are</u> all subject to causation.[1] We all had parents, who brought us up with varying degrees of loving care. We all suffered childhood traumas of varying severity, and were all exposed to role models of differing moral desirability among our peers and authority figures. We have all encountered occasional illnesses, anxieties, and provocations. These are the facts that threaten to send us uncontrollably down the "slippery slope" to moral

---

[1] I am <u>not</u> here begging the question in favor of determinism. I am only making the modest claim, which even libertarians would surely accept, that at least some of our actions are subject to causation. This leaves open the possibility that we can transcend such causes by uncaused exercises of our will, of the kind described by the "agent causation" theory (see section 4a, below).

exculpation for all actions. A "slippery slope" argument is analogous to the "domino theory" which often appears in political discussions. If a certain practice, which may in itself be unobjectionable, is allowed, it will inevitably lead to further practices which are definitely unacceptable. A familiar slippery slope argument is used in favor of a complete ban on mercy killing. While in many cases it may seem humane and desirable, it may be abused by greedy people eager to get rid of unwanted, elderly relatives. In the case under discussion, the slippery slope argument begins with the harmless-looking fact that we sometimes excuse people from responsibility. If causation negates responsibility in extreme cases, why does it not do so for our everyday actions, which are equally subject to causation?

The sciences of sociology and psychology are devoted to giving various kinds of causal explanation of our behavior. Provided that the causes involved are not the extreme ones already recognized as exculpating, it might be insisted, such explanations need not threaten our moral responsibility for everyday actions. We can, in other words, maintain a clear distinction between extreme cases where we are not to blame, and "business as usual," when we are responsible. Unfortunately, the everyday practice of modern psychiatry provides no basis for this distinction. Willard Gaylin describes two psychoanalytic principles that

are almost universally accepted among psychiatrists, even those who reject Freudianism.

> The first axiom: Every individual act of behavior is the resultant of a multitude of emotional forces and counterforces; this is the 'psychodynamic' principle. The second: These forces and counterforces are shaped by past experience; this is the principle of psychic causality.[1]

These principles require them to regard all actions, including those in which the standard excuses we have considered are absent, in the same, non-blaming light.

> The social view of behavior is essentially moralistic, an action is approved or disapproved, right or wrong, acceptable or non-acceptable. A person is guilty or innocent as more or less clearly defined in advance by law. But psychiatrically speaking, nothing is wrong - only sick. If an act is not a choice but merely the inevitable product of past experiences, a man can be no more guilty of a crime than he is guilty of an abscess.[2]

This attitude is not a mere quirk of the psychiatric profession: rather it seems essential to any scientific approach to human behavior. Any untoward behavior must be explained by the causal factors operating on the malefactor, whether they be physical brain events, provocations by other people, childhood experiences, or socio-economic background. Such explanations are the business of any of the social sciences. There are many

---

[1] Willard Gaylin, The Killing of Bonnie Garland (Penguin Books, 1983), p. 252.

[2] Ibid., p. 253. (my emphasis)

minor transgressions - shoplifting, tax evasion, sexual unfaithfulness, parking offenses - for which we plainly expect people to take responsibility. However, a consistent application of the scientific view of human behavior would require us to search for the forces which led to these actions; and blame would have no place in this scientific viewpoint. In support of this viewpoint, new correlations are constantly being found between criminal behavior and both psychological and sociological background. In fact the same kind of correlation could be found between normal, healthy behavior and antecedent causal factors, though of course such correlations are of less interest to scientists, whose goal is to reduce deviant behavior. If the slippery slope argument is effective in ruling out moral blame for misdeeds, it should also exclude moral praise for our laudable actions. Both sides of the coin of responsibility are threatened by the scientific viewpoint.

Moreover, the slippery slope argument can also be formulated without reference to scientific practice. Even our everyday beliefs are susceptible to it. It has already been pointed out that in everyday life we forgive people for minor misdeeds when we discover their causes: illness, anxiety, emotional problems, etc. More generally, we tend to be more forgiving of a person's faults when we know her well. We "make allowances" for a person's character faults

(within reason) and mitigate our blame: "Don't be too hard on her - she's just insecure/shy/impatient/hot headed, etc." On discovering that he was a murderer, how many parents would recommend that their son be executed? How many lovers would help to incriminate their loved ones? Very few, comes the obvious response, for the simple reason that these people are quite naturally biased. We need to consider why they are biased. Isn't it likely that one of the reasons is that they know their son or lover intimately? They know just how their loved one was probably thinking, the stresses and provocations to which he was subject, at the time of the crime. The parent who vehemently calls for the execution of a mass murderer would most likely moderate her demands if she knew as much about the killer as she does about her own son[1] (even though this extra information is likely to make the murderer seem even more unpleasant). Parallel to the slippery slope argument which arose from our consideration of the social sciences, can we justify the difference in our attitude toward those whose character and history (i.e., causal influences) we know well, and those who are strangers? If

---

[1] On the other hand, a parent's intimate knowledge of her child's history and character does not lead her to mitigate her praise for the latter's good deeds. A parent would scarcely write off her son's achievements with the words, "Oh, it was nothing! Given his genetic make-up and home environment, he couldn't fail!" I will not attempt to justify this lack of symmetry, since my argument deals only with our attitudes towards misdeeds.

the causation of people's actions drives out their responsibility, then shouldn't all misdeeds be forgiven? After all, a person's actions are none the less caused for the fact that, since we don't know her, we are unaware of the causes.

The "slippery slope" argument we have been drawn into may be summarized thus:

1. We are not responsible for actions caused in certain extreme "pathological" ways, e.g., chemical brain imbalances, extreme provocation.

2. No significant moral difference exists between so-called pathological, and so-called "normal" behavior, since they are equally subject to causation.

3. So we are never responsible for out actions.

The argument begins with uncontroversial cases, and shows how a consistent application of the scientific viewpoint, which is itself reflected in our everyday judgments, seems to require universal exculpation. What I consider to be the problem of free will is that this conclusion clashes with the widely held belief that we often are morally responsible for our actions (which I call the moral viewpoint). In other words, the problem is the apparent conflict between the scientific and moral viewpoints, both of which are deeply entrenched in our world view. Yet another formulation of the problem which captures my meaning is the apparent conflict between causal explanations of human behavior and moral responsibility. The advantage I claim for my slippery slope argument is

that it brings out graphically the apparent conflict between these viewpoints. Moreover, it presents the free will problem as pressing, since the conflict is between beliefs which the ordinary person can easily be persuaded that she already holds.

I am loath to give up either of the sets of beliefs contained in the moral and scientific viewpoints. The alleged incompatibility between these two sets of beliefs is asserted by the slippery slope argument. I choose to deny the conclusion of the slippery slope argument by denying premise two. Despite the fact that the scientific viewpoint can be extended to all human actions, I do not think that this leads to universal exculpation. I think that a meaningful dividing line can be drawn between responsible and nonresponsible actions even though all actions are caused. In so doing I join the ranks of the compatibilists, who believe that the scientific and moral viewpoints, and universal causation and moral responsibility, can coexist.

Those who insist that the two viewpoints cannot be reconciled are called incompatibilists. Some of them would endorse the slippery slope argument, in particular the second premise that, since all actions are caused, they are equally responsible or nonresponsible. Since, as premise one indicates, some actions are exempt from responsibility, we are never responsible. This is the view of the hard

determinist, who rejects the moral viewpoint, and regards moral responsibility as an illusion. The other kind of incompatibilist is the **libertarian**, who rejects the slippery slope argument's conclusion that we are never responsible. He agrees that **if** all actions are caused, they are all morally indistinguishable (i.e., excusable). However, he asserts that some actions are not caused, and that hence, in opposition to premise two, a line can be drawn between responsible and nonresponsible actions. Some of our actions are spontaneous exercises of our will, with no cause outside of ourselves, and it is for these actions that we are responsible. The people discussed at the beginning of the chapter are exempt from blame precisely because they, unlike most of us most of the time, are subject only to physical causation. The libertarian notion of spontaneous or self-caused actions not subject to physical causation is known as the "agent causation" theory (see section 4a below). The hard determinist rejects the moral viewpoint; the libertarian rejects the scientific viewpoint. They are agreed that universal causation rules out moral responsibility, and are hence both incompatibilists.

The compatibilist must show that, despite incompatibilist objections, the scientific and moral viewpoints are not contraries. In so doing, he must refute the second premise of the slippery slope argument by producing a criterion of responsibility which

1. Accepts the universal causation of human actions, and
2. Counts some actions as responsible, and some as excused.

2. Determinism and Indeterminism

According to my formulation of the free will problem, it is a conflict between moral responsibility and the causal explanations provided by the scientific viewpoint. The word "determinism" did not appear, although it is usually foremost in discussions of free will. The reason I did not use the term is that I believe it tends to lead free will debates into discussions of science which are not relevant to the problem.

It is understandable that the debate is so often set up as a conflict between moral responsibility and determinism. Determinism shows the scientific viewpoint in a particularly graphic way, which highlights its apparent incompatibility with responsibility. It emphasizes the force of the slippery slope argument by positing causal explanations for all actions, even if we cannot find the explanations. Here is the often quoted passage from Laplace, in which he describes determinism:

> Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situations of all the

> beings who compose it - an intelligence sufficiently vast to submit these data to analysis - it would embrace in the same formula the movements of the greatest bodies in the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes.[1]

Determinism is the view that, given its causal antecedents which may stretch back into the past, any event which occurs had to happen. Given complete knowledge of the laws of nature, and of the state of the universe at any given time, an infinite being - a "Laplacean demon" - could predict with certainty every subsequent event.[2] Not only does this image give an acute sense of the completeness of the scientific picture; but also the possibility of prediction suggests an inexhorability which further threatens moral responsibility. To a Laplacean demon, the murder committed by the victim of domestic assault described above would have come as no surprise. Even more ominously, even our trivial, inconsequential everyday actions would be equally predictable by him: what clothes I decide to wear tomorrow, what I eat for breakfast.

---

[1] Quoted by Wesley Salmon, op. cit., pp. 335-6.

[2] This is a personified version of a more rigorous definition of determinism: the conjunction of a statement of the state of the universe at any given time and a statement of the set of the laws of nature entails any statement describing the occurrence of a subsequent event. See Peter Van Inwagen, "The Incompatibility of Free Will and Determinism," Philosophical Studies, 1975, reprinted in Gary Watson (ed.) Free Will (Oxford, 1982), esp. p. 47.

Belief in determinism can safely be assumed to have been the orthodox opinion among scientists in most of the modern age. However, around the turn of the century, it was discovered that there were cases in which we were unable to predict the motion of subatomic particles.

> Everyone is agreed that it is possible to set up in the laboratory experimental systems in which events occur that we are at present completely unable to predict. This is another way of saying that we can establish no unique causal connection between the event and other events or situations, which is another way of saying that up to the present we have been able to formulate no law of nature according to which the given event follows from other things. This again may be described by saying that, as far as we can now see, the event is undetermined.[1]

Based on these experiments, physicists such as Heisenberg, Bohr, and Born set up a new branch of physics, quantum mechanics, which explicitly ruled out the possibility of unique, deterministic predictions and explanations. In defense of determinism, it might be insisted that our uncertainty is due to our limitations, and that a Laplacean demon would be able to predict with certainty even subatomic particle motions. However, part of Heisenberg's and Bohr's work is aimed at showing that it is impossible to simultaneously measure both a particle's position and its velocity, thus ruling out deterministic predictions. Moreover, von Neumann has produced a proof that it is

---

[1]Percy W. Bridgman, "Determinism in Modern Science," in Sidney Hook (ed.), Determinism and Freedom in the Age of Modern Science (Collier Books, 1961), p. 58.

impossible in principle to discover any further "concealed parameters" which would allow us to make deterministic predictions of the movements of subatomic particles.[1] Some physicists, including Einstein, have challenged these arguments, and insist on the possibility of deterministic explanations. Let us assume for the sake of argument that quantum mechanics is irreducibly indeterministic, and see whether this would be more hospitable to moral responsibility than determinism is.

Since determinism seems to conflict so drastically with responsibility, we might hope that the discovery of indeterminism is what we needed. However, indeterminism would not free us from those features of the scientific view which threaten moral responsibility. The indeterminism of quantum mechanics, no less than the determinism of classical Newtonian mechanics, posits no causes other than physical ones. The only extra "elbow room" created by indeterminism is the role of chance: these purely mechanical causes dictate the movement of particles within a certain statistical range. This is hardly the exemption from causation which the incompatibilist requires for moral responsibility.

Quantum mechanics provides statistical laws of nature. Even a Laplacean demon could not predict the exact moment

---

[1] Ibid., pp. 61-3.

at which any given uranium atom will suffer radioactive decay.[1] This might be used to argue that such processes are not subject to causation. However, in practice even supposedly deterministic processes can only be predicted probabilistically, and no one doubts that they are caused. See, for example, the discussion of paresis in footnote 1, page 2. Ah, it may be countered, but the latter predictions are only statistical because of our insufficient knowledge, whereas quantum mechanical laws are irreducibly statistical. Isn't there a qualitative difference between these two kinds of explanations, with only the kind which is in principle precisely predictable being a real causal explanation? To begin with, this is an ad hoc restriction of what counts as a causal explanation to those that can be stated as deductive arguments.[2] Moreover, exactly what element of causal connection is present in deterministic explanations, but missing in indeterministic accounts? Those who think that this element is the idea of necessary connection fall foul of David Hume's famous arguments. He argued that the only ground we can ever have for asserting causal connections is our observation of correlations between cause and effect, along with our inductively-based expectation that such

---

[1] Salmon, op. cit., p. 341.

[2] Ibid., pp. 341-3.

correlations will continue to hold.[1] Thus even our most entrenched laws of nature are "merely" statistical, in that they have been confirmed on 100 percent of the occasions we have tested them, and that we expect such success to continue. I conclude that the statistical explanations of indeterministic quantum effects are genuine causal explanations which leave no more room for moral responsibility than do deterministic explanations.[2]

In fact, compatibilists and incompatibilists alike have rejected the relevance of indeterminism to moral responsibility. Would any of the examples, discussed in the previous section, of misdeeds for which we would excuse someone, change on the assumption of indeterminism? What if the mental patient's chemical imbalance in the brain were found to involve indeterministic particle movements? Or would we change our minds if we discovered that the violent criminal's terrible home environments had influenced him via random effects in his brain? The answer to all these questions is clearly no. We are no more

---

[1] Enquiry Concerning Human Understanding, Ch. 7.

[2] I am here using "causal explanations" in the sense of explanations which refer only to physical causes. It may be objected that this notion of causation, which includes indeterministic cases, is obscure. After all, the objection continues, the constant conjunction which is the basis of Humean causation is not present here. I grant my opponent, if he so desires, this more restricted notion of causation, which excludes indeterministic explanations. As I argue in the next paragraph, even if indeterministic cases are viewed as random, chance events, moral responsibility is still threatened.

responsible for random, chance events than we are for events fully determined by antecedent causes. At this point, some compatibilists go on the offensive, and insist that deterministic causation is necessary for responsibility. They point out that it is when we are most sure of the causation of someone's actions - i.e., when they are "in character," and we see perfectly well which of his settled dispositions led him to act that way - that we are most ready to blame him.[1] Incompatibilists, for their part, are wise to abandon attempts to base moral responsibility on indeterminism.[2] Instead, those who reject determinism (i.e., the libertarians) suggest that our actions are caused by us, but are not determined by antecedent physical causes. This is the doctrine of "agent causation," described above, according to which our actions can be undetermined without being random. Both compatibilism and agent causation will be examined in depth below.

For the rest of this dissertation, I propose a convenient terminological convention. Having just argued that it is causal explanations of human behavior, and not

---

[1]See especially R.E. Hobart, "Free Will as Involving Determinism and Inconceivable Without It," Mind, 1934, reprinted in Berofsky (ed.), Free Will and Determinism (Harper and Row, 1966). The same argument recurs in most traditional compatibilist writings.

[2]Jennifer Trusted, for example, admits this in her book Free Will and Responsibility (Oxford, 1984), pp. 55-8.

just determinism, that conflict with moral responsibility, I will henceforth use the word "determinism" to mean any system of causal explanations of human behavior, including (indeterministic) statistical ones. No harm will be done as long as it is remembered that this is just an abbreviation. The major motivation behind it is to achieve uniformity with the literature, in which both types of causal explanation tend to be lumped together under "determinism." On the few occasions when it becomes necessary to distinguish the two types of causal explanation (deterministic and indeterministic), I will refer to either "Laplacean determinism" or "indeterminism."

3. Traditional Compatibilist Solutions

Compatibilists, starting with David Hume's famous treatment of the subject,[1] have insisted that the whole problem only arises due to conceptual confusions. Once we understand the meaning of terms such as "freedom," "responsibility," and "necessity" in their ordinary sense, and not in the artificial senses created by philosophers, we will realize that determinism in no way rules out moral responsibility. In so doing, compatibilists offer a criterion of responsibility which halts the slippery slope described in

---

[1] "Of Liberty and Necessity," op. cit., ch. 8. Quotations are from the Hackett (1977) edition.

section 1. Blameworthy and excusable actions can be distinguished even on the assumption of universal determinism. I will examine several doctrines shared by most traditional compatibilist accounts.

a. Causation does not entail Compulsion

A major reason for believing that determinism rules out moral responsibility is the image of causation as an agent forcing us to act against our will. In such circumstances, it seems that we cannot avoid doing what we actually do. The view of causation which encourages this image is what Hume calls "the doctrine of necessity." It is supposed that there is a necessary connection between the occurrence of a cause, and the appearance of its effect.

A major theme of Hume's Enquiry is precisely to oppose this view of causation. As we saw in the previous section, he argues that the only meaningful notion of causation is based on (1) our observation of a "constant conjunction" of causes and effects, and (2) our expectation, based on habit, that this conjunction will continue in the future. (These, for Hume, are the only grounds for our inductive predictions.) When applied to human action, this view of causation amounts to no more than the uncontroversial observation that there is a regular connection between people's character, and their actions: "The same motives

always produce the same actions."[1] Knowing a person's character, her usual desires and beliefs, we can predict with great accuracy how she will behave. Shorn of any implications of necessity, the concept of causation as a regular connection between motives and behavior poses no threat of compelling our actions. A mere description of a regularity can scarcely exert any compulsion.

Despite Hume's insistence that even in causal interactions between inanimate physical objects there is no more than a conjunction of cause and effect, a suspicion may linger that when it comes to human behavior, causation is less strict. We seem to have more "elbow room" for spontaneous, varied actions than inanimate objects. In support of this suspicion, it is pointed out that people are capable of surprising us with wildly unpredictable behavior. Hume replies with an impressive list of everyday assumptions we make based on the regularity of human actions. As for their occasional unpredictability, he points out that even complicated machines, like a watch or the human body, can behave unpredictably. In such cases, as in the case of unpredictable human behavior, rather than concluding that the effects are arbitrary and uncaused, we put the unexpected occurrence down to "the secret operation of contrary causes."[2]

---

[1] Ibid., p. 55.

[2] Ibid., p. 58.

Moritz Schlick, in his equally famous contribution to the compatibilist literature,[1] further diagnoses the confusion of causation with compulsion. All that determinism provides is laws of nature, which describe the way matter and people behave. Laws of state, on the other hand, prescribe the ways we are allowed to act, and restrict our actions by threatening penalties if we violate these laws. Only the latter kind of law can exert any compulsion on our behavior. It is not determinism, but only interventions with our behavior - whether they be man-made, like laws of state, or natural, like debilitating illnesses - which can compel our actions.[2]

b. Responsibility depends on the unimpeded following of our desires

What is needed for responsibility, according to the compatibilists, is not absence of causation, but rather absence of compulsion. There are many clear cases of compulsion in which we would excuse someone for her misdeeds. Someone breaks a window because she was pushed through it; a bank clerk is forced at gunpoint to hand over

---

[1] "When is a Man Responsible?" Problems of Ethics (Prentice-Hall, 1939), reprinted in Edwards and Pap (eds.). A Modern Introduction to Philosophy, 3rd edition (The Free Press, 1973).

[2] John Stuart Mill gives a similar Humean account of confusions arising from regarding causes as necessary, in "Of Liberty and Necessity," A System of Logic, Ch. 2, book 6 (1843), reprinted in Edwards and Pap, op. cit.

cash; a hypnotist secretly "programs" his subject to commit a murder; a party-goer has his drink "spiked," and under the influence of the drug commits a crime; an epileptic strikes someone as she is having uncontrollable spasms. The cases of diminished or no responsibility with which I began the slippery slope argument in section 1 are also cases, of varying plausibility, of different kinds of compulsion. What is absent in all these cases is the ability to act, unhindered, on what Schlick calls our "natural desires."[1] Freedom, in contrast, is the ability to act on our natural desires. Similarly, Hume defines the kind of liberty relevant to moral responsibility as "a power of acting or not acting, according to the determinations of the will."[2]

Not only does freedom require the absence of obstacles to our fulfilling our desires, it is also necessary that we be able to refrain from acting on a desire, which Hume would call "the power of not acting." Ayer asks how we would distinguish between an accountable criminal, and an uncontrollable kleptomaniac, whom we would tend to excuse.[3] Both of them are following what seem to be natural desires, but only the former could, if he wanted to, refrain from his crime. On the other hand, even if the

---

[1] Schlick, in Edwards and Pap, op. cit., p. 62.

[2] Hume, op. cit., p. 63.

[3] "Freedom and Necessity," Watson, op. cit., p. 20.

kleptomaniac wanted to refrain from stealing, his compulsion would have made him do it. Thus compulsion can be exerted not only by interferences with our desires, but by the desires themselves if we are unable to resist them.

This criterion of responsibility as the absence of compulsion, and not as the absence of causation, is supported by our actual practice in blaming people. We regard them as "the more responsible the more motives we can find for [their] conduct."[1] When we see exactly what motivated a criminal in his misdeed (i.e., what caused him to do it), and moreover that it was "in character," we hesitate the least in blaming him. It is precisely when we are puzzled by why someone committed an unexpected crime that we seek excusing kinds of compulsion which forced him to act. We would sooner conclude that he is crazy than that the crime is a wilful expression of his desires for which he is accountable. This leads to the further compatibilist doctrine, discussed in section two of this chapter, that indeterminism is of no help in establishing moral responsibility. If an action is random and we cannot discover its motivation, we are much less likely to assign blame for it than if it were a predictable action, of which we'd say "that's just what he'd do." So causation, as manifested in actions which are explicable in terms of the

---

[1]Schlick, in Edwards and Pap, op. cit., p. 66.

agent's known desires and beliefs, actually seems to be a necessary condition for, rather than a hindrance to, responsibility.[1]

c. Responsibility and punishment

A recurring feature in traditional compatibilist accounts is an attempt to rid the concept of responsibility of metaphysical, retributive notions of "absolute guilt." Instead the practice of holding people responsible is justified in a wholly utilitarian way by its beneficial consequences.[2] In particular, being responsible is equated with being a person on whom punishment will have a worthwhile effect. Punishment here includes moral criticism, as well as legal sanctions. Mill, for example, openly asserts "Responsibility means punishment," and proceeds to give a

---

[1]However, this traditional compatibilist argument has been powerfully challenged by Daniel Dennett, in "On Giving Libertarians What They Say They Want," Brainstorms (Bradford Books, 1978). Even though, as we will see, he is a compatibilist, Dennett argues that the existence of random processes underlying human behavior need not rule out moral responsibility. He points out the heuristic value of a "random consideration generator" which suggests alternative courses of action between which we may choose. The absence of Laplacean determinism would thus not make our actions chaotic and uncontrollable, and would not negate responsibility.

[2]It is interesting that Hume may be an exception among compatibilists. He seems to leave open the possibility of a retributive justification of punishment by talking of "punishment and vengeance," and by stressing the role of our moral sentiments in assigning blame. Op. cit., pp. 65-6.

"naturalistic" account of conscience, according to which the only reason we feel guilty after a misdeed is our expectation of others' hostility and punishment.[1] Whether or not one holds a utilitarian moral view, he claims, there will always be a connection between moral responsibility and amenability to effective punishment.

The most carefully-argued attempt to identify responsibility and liability to punishment is provided by Schlick.[2] He roundly dismisses any retributive justification of punishment, which aims at giving the criminal what he deserves, restoring justice, or some such ideal, as "altogether barbarous." Instead, he regards punishment as an educative measure, with purely forward-looking justifications. The criminal can be prevented by intimidation from repetitions of the crime, or better still reformed, so that he no longer even wants to act badly. At the same time, other potential criminals will be deterred from crime by seeing the consequences of being caught.

In actual practice, how do we determine whom to punish? The scientific view considered in section 1 brought out the many causal influences which led the criminal into his current state. If we regard causation as

---

[1] Mill, An Examination of Sir William Hamilton's Philosophy, reprinted in Berofsky, op. cit. See pp. 172-4.

[2] Schlick, in Edwards and Pap, op. cit., pp. 63-65.

necessary for responsibility, doesn't this commit us to the absurdity of punishing everyone and even every chemical that had a causal effect on his development? No, replies Schlick, because the people we punish are precisely those whose behavior can be modified by sanctions. Clearly to punish the criminal's father, or anyone else who influenced his early life, would fail to achieve the primary goal of reforming the wrongdoer. The standard excuses from responsibility all concern people whose future behavior is impervious to improvement by means of our sanctions. The insane person is beyond the voice of reason;[1] the person who was hypnotized very likely did not even want to commit the crime, and better consequences would follow from punishing the hypnotist; and, more controversially, it is dubious whether punishing the victim of domestic violence in "The Burning Bed" would affect her future behavior in the desired way. If she had the horrendous misfortune to find herself again in the same situation, would her previous prison term prevent her from killing again?

---

[1]However, Joel Feinberg argues convincingly that "mentally ill but rationally competent offenders" are amenable to prevention and deterrence by means of punishment. This may be true even though they don't understand their own actions, and are practically unable to control their criminal behavior, which raises serious questions about the moral permissibility of a purely utilitarian view of punishment. See "What is so Special about Mental Illness?" Doing and Deserving (Princeton, 1970), pp. 290-2.

In contrast, the people whose behavior is amenable to being influenced by our sanctions are those whose (criminal) actions were uncompelled expressions of their desires - precisely those who qualify as responsible by the criterion outlined in the previous section. Thus the "textbook" criminal has a clear motive for his crime, which is "in character" for him, and which he commits without any undue pressure from other people or extreme circumstances. Unlike many of the people described in the previous paragraph, he also has the power to resist acting on his desire. It seems likely that this kind of wrongdoer will indeed be responsive to the "behavior modification" which Schlick and his supporters regard as the goal of punishment. So Schlick has rather cleverly found support for his criterion of responsibility (i.e., responsiveness to behavior modification) by looking at our actual practice of punishment, and by interpreting this practice in a utilitarian light.

It is worth noting that the previous extended argument focuses only on the preventive and reformative goals of punishment. The equally important utilitarian goal of deterring other people would presumably be served by punishing insane people, who may not themselves be prevented or reformed by punishment. Since punishing such "innocent" people violates our moral intuitions, Schlick can hardly claim that his utilitarian view of punishment is

reflected in our actual practice of punishment. Schlick might defend his utilitarian conception by arguing that the punishment of the insane would be so arbitrary that it would weaken respect for the law, and hence reduce deterrence.[1]

We will see in the next chapter how Dennett, the compatibilist whose position I largely support, also links responsibility to a utilitarian view of punishment. My concern in Chapter V will be to examine whether someone who accepts the truth of determinism has to be committed to this view of punishment.

d. "Could have done otherwise"

One of the main motivations behind the fear that determinism rules out moral responsibility is the belief that if all of our actions have causes stretching back into the past, we cannot do otherwise than we actually do. We are merely following the only path which, given the causal conditions operating on us, is open to us. The standard compatibilist response is to bite the bullet, and to admit that given the causal influences on our behavior, which of course include our desires, it is indeed true that we must act as we do. Mill, who is particularly clear on this point, argues that

---

[1]This and other objections to a utilitarian theory of punishment is examined in some detail in Chapter V, section 2.

this is no more than a consequence of the harmless truth that: "whatever happens could not have happened otherwise unless something had taken place which was capable of preventing it."[1]

However, compatibilists do believe that there is a different sense in which it must be true that one could have done otherwise. This is what has become known as the hypothetical or conditional analysis of the phrase. In keeping with the compatibilist definition of freedom as the uncompelled, unimpeded exercise of one's desires (see section b, above) Mill states what is necessary for responsibility. It must be true that: "I could have chosen the other course if I had preferred it."[2]

Provided one has the freedom to act on one's decision, whatever it may be, then, regardless of the fact that one's decisions may themselves be caused, one has the freedom required for responsibility. Crucially, had one's decision been different, one would have been able to act on that decision, too.

This hypothetical analysis of "could" assumes that our actions and decisions are caused. However, it is sometimes objected that we have introspective evidence of more radical freedom. Thus, at the moment we act, regardless of

---

[1] Mill, in Edwards and Pap, op. cit., p. 55.

[2] Mill, in Berofsky, op. cit., p. 169. (emphasis Mill's)

the causal antecedents, we are said to feel a genuine power to perform any one of several alternative actions. Instead of walking out of the door, I feel that at any moment, however late, I could turn around, walk back indoors, and sit down. Is this not evidence that our actions are not, after all, determined? Mill replies that we do indeed have such power, but that it does not imply exemption from causation. Given the causal antecedents and states, including my desire to walk through the door, it is indeed determined that I will walk through it. In the imagined case where I turn back and re-enter the room, a crucial causal antecedent has been changed: I no longer desire to walk through the doorway. As Mill says, "When we think of ourselves hypothetically as having acted other than we did, we always suppose a difference in the antecedents."[1] Crucially, all that is required by the compatibilists for moral responsibility is being able to do otherwise *if* we had so wanted, desired, chosen, etc.

4. New Incompatibilist Challenges

The traditional compatibilist position seems to have been accepted as the

---

[1]Ibid., p. 170.

philosophical orthodoxy until the 1950s,[1] when some penetrating objections began to appear. I will consider three such groups of objections, in chronological order.

a. The causation of our desires

The compatibilists locate our freedom in our ability to act upon our desires, hence side-stepping the issue of how our desires are caused. The first wave of challenges to this view centered on persistently asking the question: What follows for moral responsibility if we take seriously the assumption that even our desires are caused?

(i) The libertarian objection

In his influential article, "Is 'Free Will' a Pseudo-Problem?"[2] C.A. Campbell first questions Schlick's claim that his analysis of freedom and responsibility reflects the everyday meanings of the terms. Schlick, it will be remembered, claimed that the free will problem was created by philosophers who perverted the ordinary sense of these expressions. Campbell gives several examples of assigning blame which seem to resist Schlick's utilitarian,

---

[1] Among the exceptions to this consensus is the view of William James, who regarded compatibilism, or "soft determinism" as he called it, as "a quagmire of evasion." See: "The Dilemma of Determinism," in Essays in Pragmatism (Hafner, 1948), p. 40.

[2] Mind, 1951, reprinted in Edwards and Pap, op. cit.

forward-looking analysis.[1] We frequently blame historical figures who are now dead for past atrocities, even though the utilitarian goal of changing their behavior is obviously no longer possible. We never hold animals morally responsible for their behavior, yet in Schlick's view it seems that we should. After all, their behavior certainly flows from their natural desires. Finally, we recognize that there are some people whom, because they are dangerous to society, we have to "punish" by incarceration, even though we believe they are sick and blameless. This is not compatible with Schlick's identification of punishment with responsibility. In effect, Campbell's criticisms apply to any utilitarian view of punishment, and call for a general defense of utilitarianism against retributive objections. A legitimate response for Schlick would be to re-interpret all of Campbell's examples in a way perfectly consistent with a utilitarian theory of punishment. Blaming dead people may have a deterrent effect on living people; we don't blame animals for the simple reason that they are unable to understand moral concepts, only punishments and rewards; and the only reason we punish anyone, not just blameless lunatics, is to protect society, and hopefully reform the criminal. So Schlick can consistently avoid having recourse to a retributive concept of "absolute guilt."

---

[1] Ibid., pp. 68-70.

However, Campbell's main disagreement with compatibilists is much harder to dismiss. He focuses on the hypothetical analysis of "could have done otherwise." He agrees that such an analysis, which requires the absence of compulsion in acting on one's desires, is indeed a necessary condition for responsibility. However, he feels that a more "reflective" conception of moral responsibility requires more than this. This freedom is of little value if our desires are themselves caused by antecedent factors, many of which are inevitably beyond our control. The question of whether we could have done otherwise is simply pushed back from our actions to our desires. It is at this level that Campbell requires a categorical interpretation of "could have done otherwise." "A condition of [a person's] moral responsibility is that he could have chosen otherwise."[1] Being able to choose otherwise, Campbell claims, requires contra-causal freedom, i.e., our choices must be exempt from physical causation. Since Campbell does believe that we are sometimes morally responsible, he, as do all libertarians, has to reject determinism.

In contrast, the compatibilists claimed that causation is necessary for responsibility, since it is above all when our actions are "in character" (i.e., caused by our desires and beliefs) that we are held responsible. Campbell concedes that on most occasions our actions do indeed flow

---

[1] Ibid., p. 75.

from our characters, which are subject to causation. In much of our everyday lives, we simply follow our strongest desire, since no moral questions need arise. However, there are occasions of "moral conflict" where what our strongest desire impels us to do conflicts with what we ought to do. On these occasions, Campbell says that we should transcend our character and use our "creative activity" in order to follow our duty.[1] These exercises of creative activity are uncaused, and it is for them (or their absence) that we may justly be held responsible. Sometimes people will be blamed even when their actions are caused by their character, since they should have exercised creative activity to resist their inclinations.

In my discussion of determinism and indeterminism, it emerged that the absence of causation would make our actions random, arbitrary events, for which we could hardly be blamed or praised. Campbell avoids this unwanted consequence by defending a version of the "agent causation" theory. Though our actions have no physical causes, they are not arbitrary, since they are self-caused exercises of our will: ". . . my choice is not just the expression of my formed character, and yet is a choice made by my self . . . the act is not an 'accident,' but is genuinely my act."[2] Campbell's view of agency commits him to the existence of a "core" to our selves which is exempt from

---

[1] Ibid., pp. 77-82.

[2] Ibid., p. 81.

causation. Though our character is indeed subject to causation, our "core," which is presumably something like a Cartesian mind or soul, is capable of spontaneously transcending this causation.

Unfortunately, Campbell's picture of the self is vulnerable to an objection made much earlier by Mill.[1] Campbell assumes a division of the self between the character, which includes our desires and inclinations, which are subject to causation; and, on the other hand, our uncaused "core," which comes into operation when there is a conflict between our desires and what we perceive to be our duty. Mill argues that when we follow our duty our behavior is just as motivated, and subject to our character, as when we follow our nonmoral desires. Libertarians make the mistake of arguing "as if conscience were not itself a desire - the desire to do right."[2] The sign of a good person is not that he resists his strongest desire (when acting on it would be morally wrong), but rather that "his desire to do right, and his aversion to doing wrong"[3] outweigh the bad desire. Commonsense intuitions can be used to support Mill's criticism. We are surely more accountable for moral decisions that flow from stable desires and dispositions than we are for rather

---

[1] An Examination of Sir William Hamilton's Philosophy, in Berofsky, op. cit.

[2] Ibid., p. 169.

[3] Ibid., p. 170.

sudden ones which oppose our character. We would be suspicious of the motives of someone who suddenly became considerate and caring, after years of egoism. This last comment is another particular application of the general compatibilist doctrine that causation can augment moral responsibility.

(ii) The Hard Determinist Objection

Hard determinists follow the libertarian in insisting that the fact that our desires are caused would undercut the compatibilist attempt to locate freedom in the uncompelled following of our desires. Whereas the libertarian tried to save moral responsibility by exempting some of our actions from causation, hard determinists bite the bullet, accept determinism and conclude that we are never morally responsible. The institutions of punishment and blame may still be retained for purely utilitarian purposes, but they should be stripped of any connotations of guilt or condemnation. While Schlick also held a utilitarian conception of punishment, he considered this conception as an explication of the concept of responsibility, and not as a replacement for it. The advantage of the hard determinist position is that it avoids the problems described above which arise for the agent causation theory. At the same time, it shares and builds on the negative part of Campbell's paper which raised a serious

problem for the compatibilist: the attack on the hypothetical analysis of "could have done otherwise."

An especially acute version of the hard determinist argument is given by John Hospers in a series of articles, of which "Free Will and Psychoanalysis"[1] is the most detailed. He draws on modern developments in sociology and psychology, in particular Freudian psychoanalysis, to show just how pervasive is the causation of human behavior, and how antithetical it seems to be towards moral responsibility.

In response to Campbell's insistence that, if determinism is true, even our desires are caused, a compatibilist might argue that our freedom is not threatened by this. While it is true that our desires have causes, it is not as though we have no control over these causes. Causation only threatens responsibility if it threatens control. People frequently do succeed in changing their desires, after deciding that some desires are unhealthy, and others are worth cultivating.[2] Thus I might be able to coach myself so successfully that I no longer even want to smoke cigarettes; or my

---

[1] Philosophy and Phenomenological Research, 1950, reprinted in Edwards and Pap, op. cit.

[2] Mill gives an especially clear description of this process of self-improvement. We can cultivate desires which promote morally good behavior, and eliminate those that tend to make us act badly, by means of self-applied punishments and rewards (which Mill calls "aversions and desires"). See Berofsky, op. cit., pp. 170-1.

self-admonitions may be so effective that I actually want to do sit-ups in the morning. I am responsible, the compatibilist might argue, provided that I am thus in control of my desires and character.

Hospers neutralizes this response to Campbell by pointing out that even this ability to shape one's character is itself subject to causation: ". . . even the degree of will power available to [someone] in shaping his habits and disciplining himself now to overcome the influence of his early environment is a factor over which he has no control."[1] It is quite true that our character is not carved in stone from the day we are born; but the extent to which we are able to improve it may well be determined by factors beyond our control. Central to Hospers' argument for the lack of control we have over our character development is Freud's notion of unconscious motivation. (This notion, stripped of Freud's more controversial claims about the ubiquity of sexual motivations, was the basis of Gaylin's claim, quoted in section 1, that all behavior should be viewed as the "resultant" of many psychological forces.) Given that many, if not all, of the motivations for our behavior will always be unknown, even to us, how can we make realistic sense of the notion of shaping one's character, and controlling one's desires? Hospers backs up this claim by

---

[1]Hospers, in Edwards and Pap, op. cit., p. 84.

giving several graphic examples of neurotic people whose behavior is hopelessly beyond their own control. They do not know the real cause of their behavior and, worse still, they invent false notions to "rationalize" their actions.

It might be objected that Hospers' examples are extreme cases of neurotic people, whom everyone agrees are not in control of their character and desires. In contrast, most people are psychologically healthy, and not subject to these hidden forces. Hospers' argument, then, leaves intact the moral responsibility of most people. What makes his argument so powerful is his reply, which extends unconscious motivation to <u>everyone</u>. He admits that "the domination by the unconscious in the case of 'normal' individuals is somewhat more benevolent" than in the case of neurotic people; however, "the unconscious is the master of every fate and the captain of every soul."[1] Those whom we like to hold most morally blameworthy - wrongdoers who are not neurotic and appear to know exactly why they are doing what they do - are thus no more responsible than mentally disturbed people. Though they do not suffer from the blatant self-deception of neurotic people, much of their reasoning takes place at the unconscious level. In particular, the "effort of will" which would be necessary for such a criminal to reform his bad habits must surely take place at the <u>conscious</u> level. If Hospers' theory is

---

[1] Ibid., p. 90.

correct, even such attempts at self-improvement are either made or left undone on the basis of subconscious processes. By the same token, those healthy people whose unconscious motivations coincide with their conscious reasoning, and who do not get involved in wrongdoing, do not deserve moral praise. They are just lucky that the causal factors which determine everyone's character, such as genetic make-up and early environment, happened to coincide in such a way that they turned out "healthy." So according to Hospers' account of motivation, all wrongdoings can be written off as ultimately due to bad luck, and all good deeds as the lucky result of factors over which we have no control. This is the very concept of moral luck against which Daniel Dennett argues. (See below, Chapter II, section B3c.)[1]

This hard determinist argument can be viewed as a more rigorous version of the slippery slope argument sketched in section 1. If we take seriously the scientific view of human action which leads us to excuse people from responsibility in certain extreme cases, it seems that we cannot justifiably hold anyone responsible, since the scientific view applies to all human actions. Determinism

---

[1]A very similar argument, though without the same detailed support from modern psychology, is given by Paul Edwards in "Hard and Soft Determinism," in Sidney Hook (ed.), op. cit. It is Edwards, rather than Hospers, whom Dennett chooses as a representative of the hard determinist view. I have concentrated on Hospers because of the greater detail in his discussion.

rules out moral responsibility because it takes away our control, not only of our character, but also of our attempts to improve our character: in short, our self-determination.

b. Incompatibility at the linguistic level

The next group of incompatibilist objections is inspired by the methods of linguistic philosophy. Rather than an attack on specific traditional compatibilist doctrines, it is an independent argument for the incompatibility of determinism and some necessary conditions of moral responsibility. Advocates of the linguistic approach propose that we solve metaphysical questions by studying the language we use to describe them: the meaning of key concepts, and the "logical properties" of the sentences we use. This shift in emphasis by philosophers from so-called ontological or metaphysical questions, to the meta-level analysis of the language used to discuss the questions, has been called "semantic ascent,"[1] and "the linguistic turn."[2] Thus Gilbert Ryle's classic contribution to the philosophy of mind[3] consisted in a detailed analysis, richly

---

[1] W.V.O. Quine, Word and Object (MIT, 1960), p. 271.

[2] The Linguistic Turn is the title of Richard Rorty's collection of essays on linguistic philosophy (Chicago, 1967).

[3] The Concept of Mind (Hutchinson, 1949).

illustrated with examples, of the everyday terms we use to describe people's thoughts, beliefs, desires, etc.: what he calls the "logical geography" of these concepts.[1] More closely related to our topic, an argument levelled at the mind-brain identity theory is based on logical differences between physiological descriptions of brain events and descriptions of so-called mental events. Brain events can be located in space, whereas it makes no sense to assign a location to thoughts, beliefs or desires.[2]

Similarly, philosophers inspired by Wittgenstein have pointed out the logical differences between different kinds of explanation of human action: causal explanations, in terms of antecedent physical states and events; and intentional explanations, in terms of people's beliefs and desires. This distinction is sometimes expressed as the reasons (intentional explanations) versus causes (physical explanations) doctrine. A well known application of this doctrine is made by Elizabeth Anscombe in her book Intention.[3] She distinguishes intentional from unintentional behavior by suggesting that only the former can be explained by reasons. This doctrine is often accompanied by the further belief that reasons cannot be

---

[1] Ibid., p. 7.

[2] This argument is stated and criticized by J.J.C. Smart in "Sensations and Brain Processes," Phil. Review 68, (1959).

[3] Blackwell, 1957.

causes. However, some proponents of the distinction insist that reasons are a type of cause. What is essential to the reasons versus causes doctrine is that there are important differences between explanations in terms of reasons and those in terms of neurophysiological causes.

Norman Malcolm, in his paper "The Conceivability of Mechanism,"[1] uses this doctrine in a way that has direct implications for the free will debate. He gives a sophisticated analysis of the logical differences between causal (neurophysiological) and intentional descriptions of human behavior. In particular, he focuses on the fact that the connections asserted in causal explanations are contingent. For example, it is a contingent fact that the stimulation of certain nerve endings causes me to yell in pain. On the other hand, he argues there is an "a priori connection between intention or purpose and behavior."[2] Part of the meaning of having a certain desire, and a belief that doing x will satisfy that desire, is that, provided one has no more powerful conflicting desires, one will indeed do x. He concludes that causal, neurophysiological descriptions, being based on contingent connections, can never be more "basic" (in a sense which he

---

[1] Phil. Review, 1968, reprinted in Watson (ed.), op. cit. Malcolm himself believes that reasons can sometimes be causes of behavior (see sections 10-15). However, the main point of his article is that if mechanism is true, intentions can never cause our actions.

[2] Ibid., section 4.

elucidates) than intentional explanations based on a priori connections. But need there be any competition between these two explanations? Why not view them as alternative but compatible accounts of human action?

Malcolm believes that the two types of description are incompatible. In particular, he holds that the truth of mechanism - a complete neurophysiological account of human actions - would rule out the applicability of intentional descriptions. The problem, he argues, is that mechanism would provide a complete account of the sufficient conditions for any action. This neurophysiological explanation leaves no room for beliefs and desires to causally influence our behavior. In fact, the truth of mechanism would entail that all of our actions would have taken place even if we hadn't had any intentions.

> If the neurophysiological theory were true, then in no cases would desires, intentions, purposes be necessary conditions of any human movements. It would never be true that a man would not have moved as he did if he had not had such and such an intention.[1]

The underlying principle used by Malcolm seems to be the assumption that the existence of one set of sufficient conditions for all human movements rules out the possibility of an alternative set.[2] (In effect, Malcolm seems to use the same principle to argue that, since it is

---

[1] Ibid., section 9.

[2] Ibid., section 9, he makes a brief attempt to justify this principle.

undeniable that we can provide complete intentional descriptions of human behavior, a mechanistic explanation is not possible. He would justify his choice of intentional over mechanistic explanations by means of his earlier argument that the former are more "basic.")

Malcolm draws the startling conclusion that if mechanism, which is consistent with determinism as I proposed to use it at the end of section 2, is true, then "people's intentions are never causal factors in behavior."[1] This bizarre claim is hard to reconcile with our everyday experience of the power of our own intentions. It would be quite extraordinary if these intentions which we daily experience were, by mysterious "countervailing factors" unknown to us, always prevented from causally affecting our behavior. It is more plausible that we have no intentions to begin with, than that our intentions always happen to be mysteriously rendered impotent. Malcolm therefore draws the further, even more startling conclusion that if mechanism is true, then we have no intentions, purposes or beliefs. This might be considered as an adequate reductio ad absurdum of determinism (mechanism), but Malcolm does not rely solely on it.

He attempts a further reductio of mechanism. Based on its apparent inconsistency with intentional explanations,

---

[1] Ibid., section 18.

this second argument questions the coherence of the doctrine. Malcolm has claimed that if mechanism is true, there can be no intentional behavior. Asserting statements, including the statement that mechanism is true, is a kind of intentional behavior. So if mechanism is true, we can make utterances which appear to be statements about it, but we can never actually assert it. "The assertion implies its own falsity by virtue of providing a counter-example to what is asserted."[1] A second paradox arises in connection with the assertion of the doctrine of mechanism. Presumably, we would like to believe that we have reasons for believing mechanism, e.g., scientific evidence, philosophical arguments, etc. However, believing something for a reason is intentional behavior, and this is precisely what Malcolm claims is ruled out by the truth of mechanism. So the second paradoxical consequence of mechanism is that one could never have a reason for believing it![2]

How do Malcolm's counterintuitive conclusions apply to the determinism/moral responsibility debate? If we have no purposes, desires, beliefs, or reasons, virtually every condition we could reasonably regard as necessary for moral responsibility is negated. We are usually excused for our unintentional behavior. If we have no purposes, all of our behavior is unintentional. Ignorance is another

---

[1] Ibid., section 22. [2] Ibid., section 25.

traditional excuse which would have to be extended to all behavior if no intentional behavior is possible, since knowledge is an intentional state. Furthermore, none of our actions could be said to be the product of our rational deliberation, thus negating what is probably the most important condition for responsibility: our control of our actions via our deliberations. Malcolm's arguments about the consequences of determinism leave no room for our rationality. Being rational is a pre-condition for all of the conditions of moral responsibility.

Because of these paradoxical consequences of determinism, and not just because it seems to rule out moral responsibility, it is clear that Malcolm's whole article can plausibly be regarded as an attempted reductio of determinism. Though he never explicitly rejects determinism, the whole thrust of his article favors a libertarian approach. Interestingly, though, at the end of the paper Malcolm admits that, despite its paradoxical consequences, there seem to be no theoretical reasons why determinism could not be verified as a scientific hypothesis. This possibility, together with the paradoxes that would ensue if it was realized, "presents a harsh, and perhaps insoluble, antinomy to human thought."[1]

Let us call the principle, endorsed by Malcolm, that determinism entails that our beliefs and desires have no

---

[1]Ibid., section 27.

causal effect on our behavior, the "determinism rules out agency principle" (DRA). This principle is still alive in today's incompatibilist literature. In her recent book Free Will and Responsibility,[1] Jennifer Trusted also tries to give reductio ad absurdum arguments against determinism. Following Malcolm extremely closely, she gives her own version of the argument that determinism is subject to the paradox that it cannot consistently be asserted.[2] Making persistent use of the DRA principle, Trusted draws further counter-intuitive consequences from determinism. If our thoughts never influence our behavior, then how can we account for the emergence of consciousness in evolutionary terms? How can a phenomenon with no causal effect have any survival value?

> . . . [conscious experiences] must be accidental features which have arisen purely by chance or as a necessary concomitant of some other characteristic which does have evolutionary significance. There is no indication as to what this other characteristic might be . . . .[3]

As its name indicates, the DRA principle has similarly unacceptable consequences for our everyday concept of a person as an agent. The notion that our beliefs and desires bring about our behavior "is based on a myth, the myth that feelings and thoughts are causally effective."[4] As with Malcolm's arguments, these bizarre conclusions leave little room for moral responsibility.

---

[1] Oxford, 1984

[2] Ibid., Chapter 10.

[3] Ibid., p. 116.

[4] Ibid., p. 132.

So far the DRA principle has been largely taken for granted, and supported only by Malcolm's arguments concerning the logical differences between physiological and intentional descriptions. Trusted offers added support for the principle in a comparison between human minds and machines, providing an interesting contrast with the views of Dennett, which are soon to be discussed.[1] If determinism is true, Trusted argues, then we are all just machines, albeit complicated ones. Our actions will in principle be just as explicable in purely physical terms as the movements of the simplest clockwork mechanism. Since we wouldn't dream of regarding machines as agents, whose behavior is caused by their conscious deliberation, we should cease to view ourselves as agents. There is a wealth of recent literature (among which Dennett's work is foremost) which suggests that a sufficiently complicated machine would indeed be viewed as an agent. Trusted's reply is that a machine can never become an agent because of "its lack of inner experiences and its incapacity to make moral judgments."[2] Curiously, however, Trusted admits the possibility that a machine might be constructed which is utterly indistinguishable from a person. In such a case, she argues, we would cease to call it a "machine" and would view it instead as an agent.[3] Thus Trusted

---

[1] Ibid., Chapter 16. [2] Ibid., p. 127.

[3] Ibid., pp. 128-9.

makes her thesis that a machine cannot be an agent immune to refutation, at the expense of its informative content. There is a suspicion that the DRA principle might be based on the same kind of stipulation: an agent is defined as a being whose actions are not determined.

Unlike Malcolm, Trusted explicitly rejects determinism, on the basis of the unpalatable consequences she draws from it. Instead, she defends a version of the agent causation theory of human action.[1]

How does the Malcolm/Trusted argument affect the traditional compatabilist position? It seems that the latter's attempt to locate responsibility in the free following of our desires would be undercut. If all intentional behavior is excluded by the truth of determinism, as the DRA principle asserts, then our desires would always be impotent, if they exist at all.

c. Determinism and Alternate Possibilities

In a series of recent articles and a book,[2] Peter Van Inwagen has insisted that we take seriously the requirement that in order to be morally responsible, it must be true that we

---

[1] Ibid., Chapters 20-22.

[2] "The Incompatibility of Free Well and Determinism," (Philosophical Studies 1975), reprinted in Watson, op. cit.; "Reply to Narveson" and "Reply to Gallois," Philosophical Studies, 1977; "Ability and Responsibility," Phil. Review, 1978; An Essay on Free Will (Oxford, 1983).

could have done otherwise than we actually did. I will concentrate on "The Incompatibility of Free Will and Determinism," the starting point of which is the principle of alternate possibilities:

> . . . almost all philosophers agree that a necessary condition for holding an agent responsible for an act is believing that the agent could have refrained from performing that act.[1]

Van Inwagen's aim is to prove that the truth of determinism entails that we could not have refrained from doing what we did. The novelty of his approach is that he gives a formal argument ("the main argument") to justify what is often stated by incompatibilists as an intuition which needs no support.

Van Inwagen defines determinism as the thesis that the conjunction of a complete description of the state of the world and a statement of all of the laws of nature entails the description of every future event.[2] J is a judge who could have saved a prisoner's life at a certain time T by raising his hand, but did not do so. P is a complete description of the world at T, while L is the (timeless) conjunction of the laws of nature. T_o is a time before J's birth, and P_o a true description of the state of the world at T_o. According to the principle of alternate possibilities, in order to be responsible, J must have been able to raise his hand at T. Here is Van Inwagen's

---

[1]Van Inwagen, in Watson, op. cit., p. 50.

[2]Ibid., p. 47.

argument that, if determinism is true, J could not have raised his hand at T. (Since a similar, equally sound, argument could be constructed for any human action, the force of the argument is that if determinism is true, no one could ever have acted differently from the way they actually did act, and hence no one is ever responsible.)

1. If determinism is true, then the conjunction of P_o and L entails P.
2. If J had raised his hand at T, then P would be false.
3. If (2) is true, then if J could have raised his hand at T, J would have rendered P false.
4. If J could have rendered P false, and if the conjunction of P_o and L entails P, then J could have rendered the conjunction of P_o and L false.
5. If J could have rendered the conjunction of P_o and L false, then J could have rendered L false.[1]
6. J could not have rendered L false.
7. If determinism is true, J could not have raised his hand at T.[2]

This formal argument can be viewed as a rigorous presentation of the following more intuitive argument. If determinism is true, then everything we do has a sufficient causal explanation in terms of the antecedent states of the world (including ourselves), and the laws of nature.

---

[1] J could not have rendered P_o false, since it is, by hypothesis, a true description of the world at T_o (which is before J's birth). No one, surely, can alter the past.

[2] Ibid., p. 52.

Crucially, many of the antecedent states of the world which cause our actions existed before our birth. If we were to act otherwise than we actually do, then we would be falsifying the sufficient causal explanation of our actual action. However, this would be impossible since (1) we cannot change causal factors which existed before our birth, and (2) we cannot falsify a law of nature. It follows that the truth of determinism really does entail that what we actually do is the only course of action open to us.

The challenge thrown out by Van Inwagen to compatibilists is to show what is wrong with his main argument. It is valid, so the only way to deny the conclusion is to show that one or more of the premises is false.

Such an attempt is made by Jan Narveson and Andre Gallois in their criticisms of Van Inwagen.[1] Though their criticisms are distinct, they both focus on the notion of "rendering false" which appears in premises (3) - (6), and argue that the incoherence of this notion makes one or more of these premises false. It is odd to say that my refraining from performing an action renders or makes

---

[1]Jan Narveson, "Compatibilism Defended," and Andre Gallois, "Van Inwagen on Free Will and Determinism," Philosophical Studies, 1977.

false statements which describe my performing it. It is even stranger to suggest that I would thereby render false statements describing the antecedent sufficient conditions for my performing the action. It is not I who would thus make these antecedent conditions fail to obtain: rather, they would have been false all along. It is precisely the fact that they were false (i.e., that other antecedent conditions did hold true) that caused me to refrain from performing the action.[1] Given this rather generous interpretation of "rendering false" (which is needed to make premises (3) and (4) true), premise (5) is no longer obviously true. It makes the assumption that it is impossible to render false conditions P_o that existed before our birth; but we can render them false, in the sense of performing actions the causal antecedents of which are incompatible with P_o.

Van Inwagen's reply is that such objections violate a vital stipulation made at the outset of his argument: P_o is a true description of the world at some time before J's birth. Given its truth, it is indeed absurd to suggest that J could do anything that would render it (P_o) false in any sense of "render."

In fact, this point of Van Inwagen's is the very same one made by some compatibilists. Advocates of contra-causal freedom sometimes refer to our feeling that

---

[1] See Narveson, op. cit., pp. 84-5.

on a given occasion we had the power to perform an alternative action. Mill's reply was that this feeling was based on the supposition that we did not want to perform the action we actually did, i.e., that there was at least one different causal antecedent. (See section 3d, above.) This feeling of freedom does not prove, then, that we could have acted differently in exactly the same circumstances. It seems to me that any attempt to deny Van Inwagen's argument by arguing that one could have done otherwise in exactly the same circumstances will prove inconsistent with determinism as he defines it.

The reference to Mill suggests another way of attacking Van Inwagen's argument. While admitting that we could never do otherwise than we actually do in exactly the same circumstances, Mill locates our responsibility in our ability to do otherwise if we so desired. Can we use this "hypothetical analysis" to deny Van Inwagen's conclusion?[1] This kind of analysis is central to the traditional compatibilist position described above. However, in his original paper Van Inwagen considers

---

[1]Van Inwagen argues that, though Narveson does not admit it, the latter's objection to premise (5) is based on a conditional analysis of "could have done otherwise." See his "Reply to Narveson," Philosophical Studies, 1977, p. 95. For reasons that will presently appear, Van Inwagen believes that his argument is immune to attacks based on such analyses.

objections based on hypothetical analyses of "could," and roundly rejects them.[1]

First, no matter how we analyze "could" in the main argument, the argument is valid. In order to deny the conclusion, a compatibilist would have to show that a hypothetical analysis of "could" would make one or more of the premises false. Certain statements to the effect that we could have done otherwise which, under a categorical interpretation of "could," are false, turn out true when the compatibilists' hypothetical analysis is applied. The best candidate for this compatibilist strategy seems to be premise (6), which denies that J could have rendered L false. Can we falsify this premise by showing that, on a conditional analysis of "could," J could have rendered L false (i.e., if he had so chosen, desired, etc.)? Certainly not. J can try and want to his heart's content, but rendering false any laws of nature is surely beyond his control.

Let us assume for the sake of argument that we could prove false one of the premises by analyzing "could" as a disguised conditional. The strength of Van Inwagen's position is that even then, he claims, he would not have to deny the soundness of his argument. In an ingenious switching of the burden of proof, he argues that if the

---

[1] "The Incompatibility of Free Will and Determinism," in Watson, op. cit., pp. 57-8.

hypothetical analysis of "could" falsifies any premise, this very fact can be viewed as evidence for the falsity of the hypothetical analysis. After all, Van Inwagen believes that all of the premises are true. He is not begging the question in favor of a categorical analysis of "could," since he devotes much of his article to defending the truth of each premise, without committing himself to any particular analysis of "could." If we accept Van Inwagen's shifting of the burden of proof, a hypothetical analysis of "could" will not in itself refute any premises. In addition, it will be necessary to refute Van Inwagen's arguments in support of at least one premise, regardless of how "could" is interpreted.

I need not devote more space to taking up Van Inwagen's stern challenge, since I accept Daniel Dennett's radical response to the main argument. Dennett would feel no need to try to refute any of the premises, since he accepts the conclusion! Given the truth of determinism, and accepting a categorical interpretation of "can," we cannot do otherwise than we actually do. He does not fear that this will remove our moral responsibility for our actions, because he makes the bold move of rejecting the principle of alternate possibilities, upon which Van Inwagen's whole approach depends. I will examine in detail Dennett's arguments for this rejection, when I turn to my exegesis of Dennett's position in the next chapter.

CHAPTER II

DENNETT'S REPLY TO THE NEW INCOMPATIBILISTS

As we saw in Chapter I, incompatibilists believe that determinism rules out several of the necessary conditions for responsibility. Each incompatibilist tends to focus on a particular necessary condition, or set of conditions: rationality, and the causal effect of our deliberations on our actions;[1] control and self-determination;[2] and the ability to do otherwise than we actually do.[3] Among the virtues of Dennett's work on free will is that, in giving a general defence of compatibilism, he argues specifically against each of these alleged incompatibilities. I will focus on a group of essays in Brainstorms,[4] and on his recent book, Elbow Room.[5] I believe that his defence of compatibilism in the latter is fundamentally correct, though in later chapters I will suggest improvements and some major changes. In view of this, I think that a

---

[1] Malcolm, Trusted

[2] Campbell, Hospers, Edwards

[3] All incompatibilists, but especially Van Inwagen.

[4] Brainstorms: Philosophical Essays on Mind and Psychology (Bradford Books, 1978).

[5] Elbow Room: The Varieties of Free Will Worth Wanting (Bradford Books, 1984).

detailed review of Dennett's arguments will be useful. Moreover, Dennett's reasoning tends to be highly condensed. My exegesis, with mostly original examples, and several references to other writers on the theory of responsibility, should be an invaluable aid to understanding him.

A. BRAINSTORMS[1] (Page numbers refer to this book)

Norman Malcolm argued, as we saw, that if our actions have prior physical causes, then it can only be an illusion that our conscious deliberation has any effect on our behavior. The real cause is the neural activity, which, together with our brain states and our sensory inputs, is sufficient to explain our behavior. Physical causation rules out the causal effect of our beliefs and desires (the DRA principle).

1. Levels of Description

Dennett's major strategy is to distinguish between different levels of description, or stances. The same action can be described from several different stances, without any incompatibility, in just the

---

[1]The most important essay for our purposes is "Mechanism and Responsibility." See also "Intentional Systems," and "Conditions of Personhood."

same way that a sentence can be translated into several different languages, without our concluding that only one or even none of the many sentences "really" expresses the proposition expressed by the original sentence. The DRA principle is misguided because it simply assumes the incompatibility of two of these stances. Let us examine Dennett's argument that the same action may consistently be described from more than one stance.

These different stances arise in our descriptions of what Dennett calls "systems": anything which behaves or acts, be it human, animal, machine, or plant. To illustrate the different stances, let us consider predictions of the next move that will be made by one's mechanical opponent in a video game, such as "Pac Man."[1] It is conceivable (though totally impractical) that a computer expert would be able to open the back of the machine, and after a laborious study of all its electrical circuits, predict exactly how the "little men" will (try to) cut you off. (Notice the unconscious introduction of an intentional term!) This is adopting the <u>physical stance</u>, and is only practicable in the case of rather simple systems; human beings are notoriously too complex for the adoption of this stance to be useful in most cases. A much more plausible method for the computer

---

[1]This is an attempt to provide relief from Dennett's frequent allusions to chess-playing computers in the articles cited above.

expert to adopt would be to use his knowledge of how the computer has been designed and programmed. For example, he may know that the "little men" in "Pac Man" are designed to speed up as the game progresses. This is adopting the design stance, and tends gradually to replace the physical stance as systems become more complex. The assumption is made both that the system follows a purpose, and also that the design is working without malfunction, both of which assumptions seem reasonable in the "Pac Man" case. The design stance is also used sometimes in describing human actions, even though, unlike the "Pac Man" case, we are aware of no designer or programmer. For example, Freudian psychologists often refer to unconscious mechanisms, such as projection, wish-fulfillment, passive aggression, etc., as the explanation of human behavior. Such explanations do not mention physical causes, but neither do they attribute rationality to our actions, and so they are good cases of using our "design" or "program" to explain our behavior. Finally, a player of "Pac Man" may find that the most effective way to play is to assume that the "little men" are rational beings "out to get him." He will put himself in their place, and ask himself what it would be rational to do, given their "belief" about his location on the screen, and their "goal" of killing him (e.g., "Aha! We've trapped him in the corner! You go that way, and I'll go this, so that we can converge on him.") When one regards a

system as rationally acting in the light of its beliefs in order to satisfy its desires one is adopting the intentional stance towards it. This is the way we normally explain and predict the actions of humans, and frequently also those of animals.

Having explained the three basic stances (with one more to come below), I need to make the crucial point that the choice between them is a pragmatic question of explanatory ease and plausibility. Frequently all three stances could conceivably be adopted, as in my extended "Pac Man" example, though one or more of the stances may become highly cumbersome to use. The intentional stance is not a special privilege that is conferred only upon humans, and it is not driven out by the possibility of a different stance being adopted.

> Whenever one can successfully adopt the intentional stance toward an object, I call that object an intentional system. The success of the stance is of course a matter settled pragmatically, without reference to whether the object really has beliefs, intentions, and so forth; so whether or not any computer can be conscious, or have thoughts, or desires, some computers undeniably are intentional systems, for they are systems whose behavior can be predicted, and most efficiently predicted, by adopting the intentional stance toward them. (p. 238)

Adopting the intentional stance toward an object presupposes not only that it has a purpose or design that is functioning effectively, (this much is needed even for the design stance), but also that this purpose or design is "optimal" (p. 5). For example, if a "Pac Man" machine were

programmed in an overly simple way, it would be easy to avoid the "little men" by spotting the regular pattern of their movements. The intentional stance would be of no use in prediction since, because of the poor design, the "little men" would not move in the most rational way, given their goal of killing you.

A final reference to the "Pac Man" example will nicely illustrate the pragmatic nature of which stance one adopts toward a system, by showing how one might change one's stance. The best policy for a new player may be to assume that the "little men" will make the most rational moves (i.e., adopting the intentional stance). However, I am told by hardened "Pac Man" players that after much practice one begins to see a pattern in the hunters' movements, and thus finds it easier to avoid them. This, of course, is a shift to the design stance, since one no longer assumes the rationality of the machine. Finally, when one puts in a quarter and the game doesn't begin, one looks for a purely mechanical fault: a coin is jammed, power has been cut, or a circuit blown, etc. (i.e., one takes the physical stance). The fact that one's choice of stance depends on one's purposes should emphasize Dennett's point that this choice is

> abstracted . . . from questions of the composition, constitution, consciousness, morality, or divinity of the entities falling under it (p. 16).

To make explicit the connection between our discussion of stances and the question of moral responsibility, we need to distinguish a fourth level of description, namely the personal stance. According to Dennett, we adopt the personal stance to a subgroup of those systems to which we adopt the intentional stance. The former stance seems to be "the annexation of moral commitment to the intentional," (p. 240), and so the intentional is a necessary condition of the moral stance. We adopt the personal stance when we view something as a moral agent, a being with rights, and morally responsible for his or her actions, i.e., as a person.

There are many examples of objects to which we could adopt the intentional, but not the personal stance. Complex computers such as "Pac Man" are best described in intentional terms, but would hardly command our moral commitment. A hunter will adopt the intentional stance toward his prey in order to predict its movements, but would scarcely be hunting it if he also regarded it as a person. Interestingly, Dennett regards even the adoption of the personal stance as a pragmatic question, "just a choice, relative to ends and desires, and not provably right or wrong" (p. 241). As a result, even when he lays out his theory of personhood most explicitly,[1] he only

---

[1]In "Conditions of Personhood," Brainstorms.

claims to have given necessary, but not sufficient, conditions.

If the intentional is a necessary condition of the personal stance, it follows that any incompatibility between the former and the physical stance will also infect attributions of moral responsibility (pp. 242-3). This section has been devoted to Dennett's arguments that there is no such incompatibility; but why is it that many people have assumed that different stances do conflict?

2. Diagnosis of Belief in Incompatibility

One reason is the fact that we most often descend from the intentional to the design and physical stances precisely when there is a failure of rationality. The intentional stance is of little use in explanation and prediction, and it seems as if rationality has been driven out by causal explanations. For example, if Dennett's famous chess-playing computer starts to make suicidal or even illegal chess moves, instead of offering an ad hoc justification for the rationality of its moves, we will drop the intentional stance. We might explain the machine's erratic behavior by reference to a loophole in a badly designed program (the design stance), or simply posit a mechanical failure (the physical stance). Similarly with human actions that lack rationality: for example, a psychologist confronted with a

neurotic person who compulsively washes his hands fifty times a day will not look for an intentional explanation. However, it must be admitted that intentional explanations of neurotic behavior are possible. Thus the psychologist could explain the compulsive handwasher's actions by simply referring to his belief that his hands are dirty. Such an explanation would, of course, be rather empty, because why he holds such a belief is precisely what puzzles us. We want, that is, an explanation from the design stance, in terms of unconscious urges and mechanisms. Such explanations differ from intentional ones in that, although they make reference to beliefs and desires, these beliefs and desires are either irrational in themselves, or else there is no rational connection between them and the subsequent action. As we saw earlier, intentional explanations of the "Pac Man" machine are of predictive value only when the machine not only has a functioning design, but also has a rational, "optimal" design. In the same way, we explain the erratic body movements of an epileptic during a fit by reference to the chemicals in her brain, adopting a purely physical stance. Dennett replies that though we often, as in the above cases, appeal to the physical and design stances when there is a failure of rationality, it by no means follows that we cannot also use these stances in "normal" cases when the intentional stance would work equally well, and hence not be "driven out":

> We associate the physical explanation with a failure of intentional explanation, and ignore the possibility that a physical explanation will go through (however superfluous, cumbersome, unfathomable) in cases where intentional explanation is proceeding smoothly (p. 243).

Another reason for belief in incompatibility is the fear that if we are physically determined, we are ultimately no more rational than primitive, dumb animals. It is held that a physical explanation of a system (like us) "puts that system in a strait-jacket, as it were, and thus denies it the flexibility required of a truly rational system" (p. 244). This is the fear of "sphexishness," which Dennett discusses in more detail in Elbow Room. I will defer my discussion of his reply to this fear until my analysis of the latter book. (See part B of this chapter, especially section 3a.)

A final reason for the belief in incompatibility is that nonintentional explanations are somehow felt to be more fundamental than intentional ones. As we have seen, when rationality is in doubt, intentional explanations are of little use and we need to rely on other stances. Moreover, a major purpose of the sciences of human behavior, especially psychology, is to go beyond intentional explanations. If a psychologist tells us that a psychopath committed a murder because he wanted to and believed that a gunshot in the head would do the trick, we feel cheated. After all, the shrink was hardly drawing on her expertise in giving such a simple answer that anyone

could have given. What we want, rather, are explanations of the mechanisms which led him to want to commit the murder, such as feelings of inadequacy or sexual frustration, or maybe just a hormone imbalance. In other words, we expect her to adopt the design or physical stances.

> In the end, we want to be able to explain the intelligence of man, or beast, in terms of his design, and this in turn in terms of the natural selection of this design; so whenever we stop in our explanations at the intentional level we have left over an unexplained instance of intelligence or rationality (p. 12).

3. A Tentative Account of Responsibility

If Dennett's account so far is correct, the existence of causal explanations of our actions is compatible with explanations in terms of our rationality: our beliefs and desires. So determinism does not rule out this necessary condition for responsibility - our rationality. In fact, it is a sure sign of a rational person that she can be caused by the adducement of reasons - rational persuasion - to adjust her behavior accordingly. We would hope, for example, to cause her not to drink a liquid by informing her that it is poisonous. However, although not all causes rule out the existence of reasons - and hence the potential for moral responsibility - surely some causes do rule out responsibility. It is now up to Dennett to provide at

least a provisional dividing line between causes which rule out responsibility, and those which allow its possibility.

To run through the kinds of cause which exclude praise and blame is a huge enterprise beyond the scope of this section, but to mention a few examples will be useful. Involuntary movements, reflex reactions, the actions of a madman, and physical coercion by a third party are all different ways in which one's normal, rational decision making process has no chance to affect one's behavior. Consider further the difference between someone shouting at you to make you jump out of fright, and the same person simply requesting you to jump. In both cases he causes you to jump, but only in the latter case does this cause operate via your rationality. Dennett's proposal is that we are responsible only for those of our actions that are caused in this latter way, in which our deliberation plays a role, unlike all the other cases in which deliberation is "short-circuited."

> The crucial point when assessing responsibility is whether or not the antecedent inputs achieve their effect as inputs of information, or by short-circuit (p. 248).

This account of the necessary conditions for responsibility is reminiscent of the classical compatibilist definition of freedom as acting in accordance with our desires. However, problem cases arise, threatening to undercut his whole account, when we consider science fiction scenarios in which clever neurosurgeons can

operate on our brains to produce new beliefs and desires.[1] When we act on these new beliefs, we will be using our rationality just as well as before, since there has been no short-circuiting of our system, and the resultant actions will be caused by our beliefs and desires. Neither has our rationality been exploited, as in deception, since once these beliefs are implanted they are left to play a causal role in our deliberation, with no interference from a third party. It seems that, according to Dennett's analysis, we would be eligible for responsibility for our actions after such surgery; but surely our actions would be the fault of the neurosurgeon, and not ours. A further unfortunate consequence of accepting such an example would be that before adopting the intentional stance towards someone, and treating him as a candidate for responsibility, we would have to be sure that none of his causally active beliefs and desires were artificially implanted.

Dennett's strategy is to question the coherence of the science fiction example, and to show that once we find a

---

[1]This is very similar to the bogeyman of the "nefarious neurosurgeon" whom Dennett considers in Elbow Room. The motivation behind the bogeyman is to suggest that, if our beliefs and desires are caused, we are just as helpless as the neurosurgeon's victims. Dennett's reply in that place, as it is throughout Elbow Room, is to show that it is a bad analogy, because of significant differences between it and the human condition. On the occasion we are considering now, however, Dennett's strategy is different. He questions the very coherence of the alleged counter-examples.

plausible description of such a case, his account of necessary conditions for moral responsibility is no longer threatened. Suppose the neurosurgeon implants in me the belief that East Lansing is in Ohio. Am I to be held responsible for any actions of mine that are caused, in part, by this belief? Very likely, as soon as I have occasion to express this new belief, my friends will incredulously inform me that East Lansing is in Michigan, and ask me whether I'm joking. Provided the surgery has left my rationality unimpaired, my instant response is likely to be amazement, and an admission that I'm wrong: "Of course it's in Michigan. What on earth made me think such a crazy thing?" This is a rather extreme example of a common occurrence: people holding contradictory beliefs without realizing it. When the contradiction is glaring, as in the above example, I think we definitely would blame the believer for failing to spot it and drop one of the beliefs. Even though my belief about East Lansing's location was artificially induced, I would certainly be blamed for my stupidity in not realizing how it conflicts with my other beliefs, especially if I have had time to spot this glaring conflict. This is assuming that I am otherwise rational, and a fit object for the intentional stance. So on this interpretation, an artificially implanted belief does not provide a counter-example to Dennett's view that our actions being caused by our beliefs

and desires is normally a ground for holding us responsible.

If, on the other hand, I obstinately cling to my eccentric belief, and refuse to be convinced by the overwhelming evidence offered by my friends - maps, addresses, witnesses - then it is clear that I am no longer rational. The intentional stance will no longer be applicable to me, and consequently I cannot be held responsible for any actions to which this belief leads me.

From this example it emerges that what determines whether I am eligible for responsibility for my actions is not the origin (i.e., artificial v. "natural") of the beliefs which contribute to them, but rather my subsequent evaluation of the beliefs in the light of my other beliefs. It is a blameworthy failure on my part if I don't realize that a belief is clearly incompatible with many other things I know, and I persist in holding it. Just how blameworthy this failure is will depend on 1) how blatant is the contradiction between the belief and my other beliefs, and 2) how rational I am in general. In the second variant of the "East Lansing" example, I would probably be totally excused on the ground of irrationality. Overall, I think it is best to view Dennett's tentative conditions of responsibility as a defeasible account. Ceteris paribus, we are responsible for those of our actions which are caused by our beliefs

and desires, as a result of our deliberation. This thesis is not refuted by the science fiction scenario, because

> one cannot directly and simply cause or implant a belief, for a belief is essentially something that has been endorsed (by commission or omission) by the agent on the basis of its conformity with the rest of his beliefs (p. 252).

Dennett's arguments so far are incomplete, and at best defend compatibilism against certain common objections. Several powerful objections remain. An opponent might argue that Dennett has constructed his machinery of stances, or levels of description, in an attempt to solve the free will problem by fiat. These stances are simply defined as being compatible with each other, whereas the whole question is whether physical, intentional and moral descriptions really are compatible. A legitimate reply on Dennett's behalf would be that he gives arguments, supported by plausible examples concerning people and machines, to show that the choice of stance really is a pragmatic issue. However, the opponent might persist, Dennett is on much weaker ground when he claims that whether we adopt the personal stance, which is our primary concern in discussions of responsibility, is also a pragmatic question. Viewing it this way would be adopting a purely utilitarian conception of responsibility ("what good consequences will praising/blaming have?") which many people would reject. (More on this point in Chapter V.)

This leads to a more fundamental objection. Dennett's opponent may concede that he has proven the compatibility

of purely physical and intentional descriptions of human behavior. However, being rational (i.e., the object of intentional descriptions) is only a necessary condition for moral responsibility. For example, intelligent chimpanzees are almost certainly capable of rational deliberation;[1] yet would scarcely be held morally responsible for their actions. Dennett needs to show that determinism is compatible with those further conditions - whatever they may be - which together are sufficient for moral responsibility.

B. ELBOW ROOM (page numbers refer to this book)

1. Overview

In Elbow Room Dennett takes on this more ambitious task of accounting for a wider variety of necessary conditions for responsibility. He aims to show how each condition can be perfectly accommodated within his deterministic, naturalistic picture of human behavior. The book shares with Brainstorms a naturalistic account of consciousness, which draws on Dennett's functionalist theory of mind. He is especially fond of "evolutionary" accounts, starting with the most rudimentary life forms, which we wouldn't dream of calling conscious, much less

---

[1] For a good example, see Peter Singer, Practical Ethics (Cambridge, 1979), p. 96.

responsible. He proceeds to show how, by a series of purely deterministic increases in complexity, we gradually arrive at conscious, rational, and responsible beings: us. Another feature common to both books is his use of computer and other machine analogies. His plausible examples of machines which are best viewed as rational deliberators help further allay concerns that deterministic beings cannot be rational, conscious or responsible.

The novelty of Dennett's approach, as compared to other compatibilist accounts, is that he does not only directly attack specific incompatibilist doctrines. Instead, adopting a soothing tone, he tries to undermine the fears on which he claims incompatibilism is based. He wants us to realize that the demands for exemption from physical causation as a requirement for responsibility are based on philosophical errors. Indeed, the whole free will problem

> will turn out to be a misnamed and misbegotten amalgam of overhasty problem posing and self-induced panic, the false pretext for much otherwise unmotivated system building and metaphysical tinkering (p. 6).

He wants to show us that The Varieties of Free Will Worth Wanting (the book's subtitle) are possible and indeed actual in our deterministic world.

Dennett first describes and diagnoses the "bugbears" or bad analogies which, he thinks, motivate the fear that determinism drives out responsibility (Chapter 1). He substantiates his claim that they are indeed bugbears by

means of a naturalistic account of human action which, though it assumes the truth of determinism, leaves ample "elbow room" for moral responsibility (Chapters 2-5). He runs through several necessary conditions for responsibility, each of which, it has been claimed by various philosophers, is negated by determinism. One by one, he shows how each is compatible with a purely deterministic view of the world. The conditions are rationality, control, self-determination, and effectiveness of deliberation. Chapter 6 involves a radical attack on the normally unquestioned assumption that in order for someone to be responsible, it must in some sense be true that she "could have done otherwise." Dennett argues that this question is irrelevant to moral responsibility, and that our only concern in appraising people's wrongdoings is to assess their character, or to predict and control their future actions. Dennett's concern with such forward-looking questions becomes even clearer in Chapter 7, in which he gives an openly utilitarian justification of our practices of holding people responsible, and punishing them. (The relation between compatibilism and a utilitarian view of punishment will be the topic of Chapter V of this dissertation.) It is worth noting that Dennett's arguments are more speculative and tentative in the final chapter. This suggests that he is more comfortable with the therapeutic, Wittgensteinian role of removing

confusions, than with setting out a positive theory of responsibility.

In support of this observation, nowhere does Dennett claim that he has given a complete list of the conditions that are sufficient for responsibility. Instead he has given a detailed defense of the compatibility of determinism with those necessary conditions which are commonly said to conflict with it. Consequently, I think his position is best viewed as shifting the burden of proof to the incompatibilist. It's as if he's saying: "I've shown how all the alleged conflicts between determinism and the necessary conditions of responsibility are only illusory. Now you show me any further grounds for believing in incompatibilism." This interpretation is supported by his closing statement. He expects and welcomes new incompatibilist arguments (p. 171), but maintains the therapeutic, problem-dissolving[1] tone of the whole book, by requiring us to ask:

> whether you have any clearly statable reason to hope you have that variety [of free will], any reason to fear that you might not (p. 172).

---

[1] See Ronald Suter, Are You Moral? (University Press of America, 1984), pp. 21-4. His discussion pertains to moral debates, but it is equally applicable to nonmoral questions. He distinguishes between solving a problem, which means taking a stand on it, and dissolving it, which involves rejecting one or more of the premises which create it. For instance, one might view the problem as being created by offering a false bifurcation.

This shift in the burden of proof may be viewed as a moderate exercise of William James' "Will to Believe."[1] We can safely assume that all philosophers would prefer to find a solution that will leave our free will and responsibility intact (p. 168). If this solution is one which also does justice to the increasingly prevalent scientific view of people as subject to causation, then so much the better. Of course, such a consideration "creates the suspicion of wishful thinking" (p. 169), and cannot be used as evidence for the truth of compatibilism. However, given that Dennett's arguments for compatibilism are compelling, such pragmatic reasons do seem to justify our believing in compatibilism until further notice, pending new and more powerful incompatibilist arguments.[2]

What follows is a chapter-by-chapter exegesis.

2. The Bugbears (Chapter 1)

Dennett describes "with deliberate disrespect" the metaphors which, he claims, "have done most of the work behind the scenes in propelling the free will problem" (p. 7). He uses the useful term "intuition pump" to refer to the analogies and thought experiments which, he

---

[1] Essays in Pragmatism (Hafner, 1948), pp. 88-109.

[2] A further pragmatic reason for believing in free will, which Dennett's position enables us to do, emerges in his Chapter 5. If we don't believe in our freedom, our deliberation may be hindered by feelings of apathy and resignation, and our actual freedom may become restricted.

claims, play a central role in philosophical reasoning. He examines a group of intuition pumps which were either bad to begin with, or else have been abused by being taken too literally or applied too broadly. The effect of all of the "bugbears" is to make us fear that the truth of determinism would entail the absence of various vital elements of freedom. In subsequent chapters he tries to allay these fears.

The first group of intuition pumps centers on fears that determinism would take away our control of our own lives (pp. 7-10). Maybe we are trapped within the chain of caused events, unable to escape the clutches of the "invisible jailer." Similarly, many other personifications of the forces which cause our actions are made, drawing on the very reasonable fear we have of our lives being controlled by other agents. This malevolent agent may either cause us to act just as he wishes (e.g., "The Nefarious Neurosurgeon"); or else take advantage of determinism and his knowledge by predicting our every move, and waiting for us to "play into his hands" ("The Malevolent Mindreader").

A particularly important intuition pump, which Dennett opposes throughout the book, is provided by the concept of Sphexishness (pp. 10-13). This is named after the sphex bee, a creature which shows the severe limitations of its rationality by its inflexible behavior. It acts like a

dumb machine, determined to carry out its clockwork mechanism come what may, no matter how inappropriate it is to the circumstances. A good example of sphexish behavior is the mouse who is unable to distinguish cheese in a pantry from cheese on a mousetrap. The cheese in the pantry tasted great, so why not eat this, too . . .? The fear aroused by this image is that if we too are deterministic beings (it seems clear that the sphex is), our actions too may be ultimately based on an inflexible clockwork program, which just happens to be more complex than that of the bees. To a superior intelligence, we may already appear sphexish, and even to we mortals our peers' behavior can sometimes appear this way. Consider the "creature of habit" who is stumped by a novel situation, and reacts by following the hopelessly inappropriate old routine. A graphic example of human sphexishness occurs in the movie "Being There." Peter Sellers plays Chauncey Gardener, whose entire concept of the outside world is based on what he has seen on TV. On a rare venture into the outside world, he is confronted by a gang of thugs, and to try to escape them, he changes the channel on his automatic TV control box! The fear of sphexishness casts doubts on our rationality, an important precondition of moral responsibility.

Another group of analogies focuses on the fear that, if determinism is true, our very existence as agents seems to

be swallowed up by a chain of physical events.[1] It is as if our bodies will continue to function like clockwork mechanisms, regardless of whether we want them to ("The Disappearing Self," p. 13). Even more ominously, our realization of this implication of determinism may lead us to apathy and resignation. Hence it may be in our interest to take the apparently irrational course of discouraging belief in "The Dread Secret" (p. 14), even though we have good reason to think it true. The fear that our deliberations are useless charades with no causal effect on our behavior is nicely captured by Dennett's metaphor of body English (pp. 15-16). These are the "squirms and twists and leans" that are made by golfers and bowlers who are "willing" the ball to reach its target. Of course, this performance has only superstitious value. Just as the ball has already left the bowler's hand, so it seems that the causal antecedents have already determined our behavior, making our deliberations just as superfluous as the bowler's body English.

It is in Chapters 2-5 that Dennett argues in detail why these fearful intuition pumps are not applicable to the human condition. However, he cannot resist throwing in a metaphor of his own to counteract the ones he opposes. The

---

[1] Thomas Nagel describes the danger that causal explanations of human behavior will reduce our "area of genuine agency" to "an extensionless point." (Mortal Questions, Cambridge, 1979, p. 35). More on this in Chapter IV.

justification for this tactic, which he also uses on other occasions, is that incompatibilist intuition pumps are often stated without any justification of why they fit the human condition. Given this carefree use of analogies, Dennett feels he has the right to use his own, even though it has a similar lack of detailed argument. Thus he describes the notion of follow through, common to golf, baseball, tennis and other sports. Once the ball has left the club, bat or racket, continuing one's swing can no longer affect the ball's trajectory. However, the extra effort required for a swing "through" the ball can have an important effect on one's swing at the moment of impact. This analogy has the salutary effect of replacing an ominous metaphor of useless activity by a comforting, familiar one justifying this activity. Still, I question its value. It suggests that human deliberation is somehow "after the fact," and is merely a means of deflecting the causal processes which are already under way regardless of our deliberation. Given the compatibilist drift of Dennett's whole book, a more suitable metaphor would be one in which deliberation is viewed as part of the causal antecedents of action, rather than a slight last minute interference. However, Dennett would doubtless reply that his only goal was to show how easy it is to provide analogies to favor one's position, and would not want to defend the accuracy of his metaphor.

3. The Naturalistic Account

a. Rationality (Chapter 2)

One of the least controversial necessary conditions for responsibility is rationality. Both of Aristotle's conditions for voluntariness, knowledge and control (i.e., lack of compulsion)[1] would be threatened by failures of reasoning powers. In "Mechanism and Responsibility," discussed above, Dennett argued convincingly that human rationality can and does exist in a purely deterministic world. In this chapter, he further examines the fear that determinism drives out rationality. His arguments are especially relevant to the Malcolm/Trusted argument for the incompatibility of physical causation and causally effective reasoning. Indeed, this chapter can be viewed as an important part of Dennett's reply to the DRA principle. He focuses on the image of the sphex, and the fear that ultimately our behavior may be just as inflexible and beyond the control of reason.

Even though the sphex's behavior does have purposes (e.g., building hives, collecting honey), we do not attribute it much reasoning power. We are all too aware of its limited repertoire of responses and view it as little more than a machine with only two or three different

---

[1] Nichomachean Ethics.

settings. We explain its behavior as the effect of a few biological urges. We normally explain human behavior, on the other hand, in terms of our reasons and goals. The threat apparently posed by determinism is that our actions, too, would have to be viewed not as the result of rational deliberation, but as the outputs of machines, albeit more complicated ones. This view, of course, makes the familiar assumption that reasons and causes of human behavior are incompatible (pp. 27-28). Dennett rejects this belief. He reiterates his functionalist theory of mind, according to which any system of sufficient complexity and adaptability, be it animate or artificially created, is conscious and may be rational.

Dennett tries to alleviate fears of sphexishness by sketching an evolutionary account of the development from single-celled beings to rational humans. Surprisingly, the sphex is rather high up the evolutionary ladder, having developed at least a few responses to its environment. The importance of this evolutionary account is that it shows that a determininistic being is not automatically prevented from being rational. Presumably, in our gradual evolution from simple, sphexish beings, we did not suddenly become exempt from physical causation; and no one doubts our status as rational beings. Dennett insists that it is not whether or not our actions are caused that dictates whether

we are rational; rather it is whether the causal mechanism is sufficiently complex and versatile.[1]

But how can a deterministic mechanism with no understanding (a "syntactic engine") develop into a conscious reason considerer (a "semantic engine") like us (p. 28)? He describes our evolutionary development in terms of improved responses to our environment, each improvement explained by its survival value. The final stage which creates the potential for the widest flexibility of responses is the ability to reflect on one's own thought processes.[2] Computers are being designed with this ability to learn and monitor themselves, strengthening the view that we too are rational even if our actions are determined. After all, we are still far more complex and versatile than any existing computer.

However, this last observation may be used in an attempt to drive a qualitative wedge between physical systems and us. It is insisted that any physical system will eventually reveal its limitations by confronting a new situation which it has not been programmed - either biologically or by the machine's designer - to deal with.

---

[1]Why is it so commonly assumed that causation drives out the rationality of our behavior? Dennett conjectures that it is because of a preoccupation with over-simple examples of causation, such as Hume's billiard balls, which are even less adaptable than the sphex. We need to realize that the causation of human action is much more subtle and complex. See Elbow Room, pp. 31-34.

[2]This is what Dennett calls "going meta-" (p. 29).

When it reacts inappropriately, we will exclaim, "Aha! So the whole time it only seemed to be rational, whereas it was really just a blind mechanism." We, on the other hand, suffer from no such limitations, and hence may justly be viewed as rational.[1] Dennett's response is to deny the assertion that we are not subject to limitations. He points out the permanent possibility of our confronting a superior being, from whose higher, more rational vantage point our behavior may seem sphexish (p. 46-48). This may be especially true when we indulge in ethical and meta-ethical reflection, when we frequently arrive at what appear to be bedrock intuitions which we find difficult to challenge or justify. However, Dennett's crucial point is that this possibility in no way impugns our rationality. Any physical being, whether made of computer hardware or living tissue, is of finite complexity. To demand a "friction free" (p. 29) rationality which can adapt to any situation is to ask for the impossible. Rather than making these "absolutist" (and unfulfillable) demands, Dennett suggests that we should be happy as we are. The self-monitoring meta-thinking rationality we already possess is the only kind worth wanting.

---

[1]Dennett refers to Lucas' attempt to use Godel's incompleteness theorem to prove this point. It is true that there will always be at least one statement whose truth a machine will never be able to prove. However, Lucas failed to see that exactly the same limitation applies to any physical system, including us. See "The Abilities of Men and Machines," Brainstorms, Chapter 13.

> The perfect Kantian will, which would be able to respond with perfect fidelity to all good reasons, is a physical impossibility; neither determinism nor indeterminism could accommodate it. But that does not leave us in 'sphex's' predicament. We are not infinitely but only extraordinarily sensitive and versatile considerers of reasons (p. 49).

b. Control (Chapter 3)

The fear that determinism takes away from us control of our own actions is common to most incompatibilists. The Malcolm/Trusted view fears the loss of control which, they think, would follow from the fact that physical events, and not our rational deliberation, cause our behavior. The Campbell/Hospers/Edwards view fears that if our very desires and character traits are caused, then our control over our own actions is illusory. (This was the basis of their attack on the traditional compatibilist view.) The intuition pumps, described in Section B2 above, which compare determinism with our being controlled by other agents, are also intended to provoke this fear. In this chapter, Dennett tries to undercut this fear by arguing that causation does not constitute control, despite the fact that the two words are often used as if they were synonymous.[1] His argument is akin to the traditional compatibilist accounts which equate freedom (and hence control of our destiny) with absence of

---

[1] cf. "Causation does not exert compulsion," (Chapter I, section 3a).

compulsion, rather than absence of causation. However, he goes beyond these accounts by giving a detailed analysis of the concepts of control and causation.

He first gives an analysis of the concept of control, showing how it is independent of causation. Let us take a simple, everyday example of control: driving a car. When the steering or brakes fail, we lose control; yet even while the car is swerving wildly across the highway, it is still subject to the laws of motion and gravity. Its movements are still caused. Conversely, when the car is working perfectly, and we are in control of it, it is still just as subject to the laws of nature as when it is out of control. These laws of nature are not, of course, up to us to change: they are a "given." The reason why we retain control of a car, even when many of the causal influences on its movement are beyond our control, is because we are able to anticipate and compensate for these causes. Gravity is so constant that it poses no threat to our control. Ice, on the other hand, can be hard to anticipate, and even if we do so, it may be too late, and our efforts to compensate for its effect on the wheels' traction may be impotent. Our control depends on our anticipation of and compensation for causes, not on the absence of causes (pp. 52-57).

This is all very well, it might be replied; but what about our control of ourselves? Isn't this what is

threatened by determinism, since our own attempts to control other events and objects may themselves be caused by external factors? First, there is the fear that nature controls us by natural selection. Dennett does not deny that we are caused by nature to be the way we are. This is indeed the thrust of his evolutionary account of rationality sketched in the previous section. However, since we learn to adapt to our environment in the most appropriate way, and to change our policies and even our functions accordingly, we are best viewed as controlling ourselves. The environment tends "to design creatures that have in themselves a benign tendency to make the right discriminations - for themselves" (p. 59). Dennett blames this fear that nature controls us on a preoccupation with over-simple (and hence rare) examples of causation. Scientists perform "controlled" experiments in which all variables except one are held constant, in order to isolate and measure the one causal factor considered most crucial. This is often the factor on which it is most useful for us to concentrate, since it is the easiest one to manipulate[1] (again fuelling the temptation to equate causation with control). However, most cases of causation

---

[1]This is what Joel Feinberg calls making causal citations from "the engineering standpoint," using "the handle criterion" of causation. This is one among several standpoints, the choice between which is pragmatic. See "Sua Culpa," in Doing and Deserving (Princeton, 1970), esp. pp. 204-5.

are far more complex, providing very few opportunities for control, yet remain "equally good cases of causation" (p. 60).

Second, the fear that determinism would take away our control is fuelled by analogies in which other agents cause us to act. A hypnotist or a neurosurgeon could, in these examples, cause us to do whatever they want. We are clearly "under their control." Following a line of argument suggested in Brainstorms (discussed in Section A3 above), Dennett takes the sting out of these analogies by giving his own examples in which our behavior is caused by other people benignly. Who could possibly object to being caused to act by being presented with full information and good reasons? Being amenable to such causation is surely the mark of a rational person, and something we all desire. It is not the fact that our actions were caused that made them out of our control in the first, malevolent examples. Rather, it is the fact that our reasoning processes were by-passed and played no causal role in our decisions and actions (pp. 64-66).

Joel Feinberg discusses some interesting cases of causation of our actions intermediate between the two types we have considered.[1] He discusses whether these actions would be called voluntary, but for our purposes we can

---

[1] "Causing Voluntary Actions," Ibid., Chapter 7, esp. pp. 167-73.

consider whether the agent is in control of his actions. In these cases the "causer" of our actions is malevolent, but he achieves his goal without by-passing the agent's rationality. Instead, the manipulator exploits his victim's rationality by "triggering off" certain dispositions to act which would normally remain latent. Such is the unscrupulous manager who misleads his boxer into believing he has a chance of success, and tantalizes him with the prospect of riches and escape from the ghetto. A case in which we would be less hesitant in saying that the manipulator controls the agent is Iago's goading of Othello, since Iago in part creates the disposition to jealousy in Othello. In general, we say that the would-be exploiter controls his victim's actions when the disposition which he exploits (e.g., desire to earn easy money by a dangerous means; jealousy) was unlikely to be triggered without his persuasion. If, on the other hand, it is likely that routine events would eventually have triggered the disposition, we are more likely to say that the agent was in control (and should have resisted).[1] This distinction becomes particularly acute and important when the charge of police entrapment is made. Did the police officers "lead on" the alleged

---

[1]It should be noted that my tentative examples of control all involve a malevolent exploiter. If the disposition triggered is beneficial to the agent, we are likely to view the persuasion as advice, rather than as control.

criminal, or were they merely assisting him in what he would have done anyway?[1] These examples further illustrate Dennett's point that the mere fact that an action is caused has no bearing on whether it was within the agent's control. The question of control is more complex.

c. Self-Determination (Chapter 4)

The incompatibilists mentioned above who are afraid of loss of control might persist that Dennett's arguments so far have failed to address a deeper problem. Though he may have shown that we can control our everyday actions, doesn't determinism restrict our control of ourselves? Given our genetic endowments, our strength or weakness of will, and our environment, do we really have the power to improve our character if we want to? Dennett's strategy is to extend his naturalistic account of mankind to include self-determination (or "self-definition," as he calls it), in a way which is perfectly compatible with the causation of our actions. However, this account runs afoul of a problem it was designed to avoid: moral luck. He closes the chapter with a lengthy account of why moral luck does not rule out self-determination and responsibility.

---

[1] Ibid., pp. 173-76.

A major motivation for fearing that determinism would prevent self-determination is the belief that it would leave no place for the self at all. It seems as if human actions can be explained solely in terms of sensory inputs, internal chemical reactions, including brain events, and then our overt behavior. Where is there any room in this account for a rational, responsible deliberator? (Hence the section title, "The Problem of the Disappearing Self.") "If we are mere conduits of causation, it seems, we cannot also be agents" (p. 76).[1] Dennett claims that this belief "leads almost irresistibly to the doctrine of 'agent causation.'" This doctrine carves out and protects a place for our agency as self-determiners by exempting our "self" from the causal chain. The self is an "unmoved mover" which initiates our actions without itself being subject to causation. I think that Dennett is inaccurate in suggesting that the fear of the self being swallowed up in the causal chain in itself leads to the doctrine of agent causation. It only does so in conjunction with the further belief that absence of causation would be even less hospitable to freedom.[2] Agent causation has the best of

---

[1]This can be viewed as an alternative statement of the "determinism rules out agency" (DRA) principle, discussed in Chapter I, section 4b.

[2]See R.E. Hobart, "Free Will as Involving Determinism and as Inconceivable Without It," Mind, 1934, and also Schlick, op. cit. Though they both conclude that physical causation is indeed necessary for responsibility - a belief

both worlds: exemption from causation without randomness.[1]

Why has the agency theory been so popular? The intuition that our actions have no physical causes is explained by "the magnification of effects by the nervous system" (p. 76). Tiny stimulations of our sense organs lead to our overt actions. Surely this extra energy can only be accounted for by the existence of an uncaused self which initiates actions. Allied to this is the intuition that this uncaused self must be located somewhere within us: what Dennett calls "Central Headquarters," (pp. 78-80), analogous to Ryle's "Ghost in the Machine."[2] However, our attempts to locate Central Headquarters by introspection only manage to stumble upon a series of thoughts. We do not directly intuit any initiating self.[3] Since we cannot find by introspection this alleged locus of our actions, we "exploit the cognitive

---

which agency theorists explicitly reject - their thesis that mere randomness does not allow for responsibility is part of the motivation for the agency theory. An important theme in Dennett's work, however, is that randomness does not rule out rationality and responsibility. He describes the value of random processes for our decision-making procedure, pp. 66-9, and elsewhere.

[1] See my discussion in Chapter I of Campbell and Trusted.

[2] Gilbert Ryle, The Concept of Mind (Hutchinson, 1949), pp. 11-24.

[3] This is no more than a re-statement, in more modern terms, of Hume's famous argument against mental substance, quoted by Dennett in Elbow Room, p. 74.

vacuum," and in its place posit "a rather magical and mysterious entity, the unmoved mover, the active self" (p. 79).

Dennett leaves room for the self in his naturalistic account of the self-determination that occurs within a single person's deterministic life. The self is to be found in the whole process of self-definition, and it is a mistake to try to locate it at any particular time or place within this process. A child, like many animals, is only able to follow its inclinations, whereas as we develop into rational adults, we become capable of meta-level decisions. Thus we may decide that, despite the pleasure they bring, smoking and drinking are too harmful to our long-term health for us to continue them. At a higher level of reflection, we may question the moral acceptability of our character traits, occupation, or even our lifestyle, and take steps to change them. Because we are finite, it will always be possible (but time will not permit us) to ascend to higher and higher meta-levels of self-appraisal. However, as we saw in our discussion of rationality, we are already amply capable of enough meta-criticism to allay fears of sphexishness. This process should also reassure us that we have a self, and that we are in control of this self. Furthermore, it is rational to limit our self-reflection, for fear that we may hesitate too long, or end up being incapable of deciding

either way (pp. 86-87). The kind of self determination that is especially important for moral responsibility is debate about our ultimate values, as exemplified by our solutions to moral dilemmas.[1] These decisions are importantly different from other kinds of choice in which, if we make a mistake, we may be able to pass it off as a human error. Crucially, in these meta-debates about our ultimate values, what counts as an error is itself determined by our debate. We can hardly blame our errors or misdeeds on factors beyond our control, since we ourselves create the criteria of goodness/badness. As Dennett says, "we create our values while creating ourselves" (p. 90). I discuss this notion of creating our own values in more depth in Chapter III.

However, this whole account of responsibility for self-determination is threatened by the relentless hard determinist challenge thrown out by Hospers, discussed in Chapter I, section 4a(ii) above. Dennett considers a similar argument by Paul Edwards. Isn't our ability to "rise above" our character and take responsibility for it just as determined by factors beyond our control as the fact that we begin with the character we happen to inherit?

---

[1] Dennett discusses at some length Charles Taylor's "Responsibility for Self," reprinted in Gary Watson (ed.), Free Will (OUP, 1982). The notion of "radical re-evaluation" comes from Taylor, who discusses Sartre's famous example about the Resistance fighter torn between obligations to his mother and to his country. I will discuss Taylor in depth in Chapter III.

The person who rises from the ghetto was lucky: he had the right genes, and the right home environment to give him the strength to improve his lot. His neighbor who turned to crime was just unlucky, and was born without enough will-power for effective self-improvement. The existence of moral luck seems to sabotage any attempt to locate moral responsibility in our ability to improve or change ourselves.

> How could any deterministic process of 'character transformation' beginning with a being that was not responsible for any of its 'decisions' ever yield a being who was not only responsible for its decisions, but responsible for having the sort of character that would make those decisions? For that matter, how could an indeterminate process from the same starting point yield anything better? (p. 84)

To his credit, Dennett attacks this argument head-on, and offers a first pass at a reply (pp. 84-85). The argument that nonresponsible beings can never become responsible ones seems similar to obviously fallacious ones. At one point in history, there were no mammals on earth. Does it follow that no mammals could ever evolve out of non-mammalian material? Clearly not, since we exist. Moreover, Edwards' argument "presupposes that one could not take full responsibility for something unless it was entirely of one's making" (p. 85). Only God would thus qualify for responsibility. In everyday life, it seems clear that we don't hesitate to assign responsibility to people for their creation of certain crucial necessary

conditions for an event. Presumably the bank robber didn't cause the employees to be at the bank, yet without them the hold-up could not have happened. However, Dennett admits that these responses seem to leave untouched the main thrust of the hard determinist challenge. Granted, gradual acquisition of responsibility is possible; and granted, we assign responsibility provided some of the causes of the harm/benefit were within the agent's control. But Edwards' argument maintains that all of our actions - even those character changes by which we might be hoped to gradually acquire responsibility - are ultimately due to genetic and environmental factors over which we have no control. It is this stubborn residue of moral luck to which Dennett devotes the rest of the chapter, in what I consider to be the book's most important section.

Dennett argues that the existence of luck does not rule out the possibility of responsibility. He gives a detailed analysis of the concepts of luck and skill, and how they relate to responsibility. It is one of the most difficult passages in the book, and it will be helpful to start with an analogy with a more familiar compatibilist argument. As we saw in Chapter I, traditional compatibilists argue that incompatibilists require too much for freedom and moral responsibility. Instead of requiring absence of causation, they should be satisfied with absence of compulsion. The incompatibilist has made the mistake of believing that

causation entails compulsion. In this light, let us approach Dennett's argument. It is no doubt true that, if all of our actions are due to either good or bad luck, then we cannot be held responsible. From this it follows that some exemption from moral luck is needed for responsibility. However, those who fear the implications of moral luck go further, and require absence of causation for moral responsibility. The crucial assumption they make is that, if all events are caused, then all of our actions are due to either good or bad luck. (Either way, we can't be responsible.) This assumption is precisely what Dennett challenges: "there is a tendency to treat 'lucky' and 'unlucky' as complementary and exhaustive, leaving no room for skill at all" (p. 97). Dennett, in contrast, insists that our skill, and not just luck, influences our actions. But does anyone really hold such a paradoxical view about luck (i.e., the one Dennett criticizes)? Moreover, is it necessarily wrong? Let us examine Dennett's analysis of luck and skill, which makes good use of sports examples.

Skill is a "projectible" property, in that the fact that someone exhibits it on some occasions is a reason for expecting more of the same in the future. Luck is certainly not projectible, since it's not an enduring property of people. Individual shots in basketball may be lucky, but "luck tends to average out" (p. 94). The more skillful a player is, the less he needs luck. Moreover, we

are less likely to attribute his string of successful shots to luck, for, being a star, "the threshold for what counts as luck is considerably higher" (p. 95). So on many occasions bad luck may be overcome by skill, and luck is of little significance. Dennett suggests that, in life just as much as in sports, the lucky breaks and the unlucky setbacks will tend to even out, and on the whole it is people's talents that control their actions.

However, this does not apply to differences between people's inborn character traits and abilities, which are a matter of what Nagel calls "constitutive luck."[1] Averaging out over time does not seem applicable here. It seems more like a once only dealing out of cards, and the hardened criminal who has been involved in crime since his rough, traumatic childhood was simply dealt a bad hand. Can we blame him for not overcoming such odds? Dennett replies with another sporting analogy. A ten-yard handicap would be a terrible disadvantage in a sprint; but in a marathon, it will become negligible, among all the other ups and downs in the long race. Dennett thinks that our moral development is more like a marathon: there is ample opportunity to compensate for a poor start. Moreover, our moral development is not a race at all, but rather progress toward a minimal plateau shared by everyone capable of

---

[1]Thomas Nagel, "Moral Luck," reprinted in Watson (ed.), Free Will. See p. 177.

self-determination. This moral community "is held to be composed of individuals of roughly equal talents, insofar as the demands of such [a community] are concerned" (p. 96).

It might be insisted, in opposition, that the central role of luck in our progress has not been eliminated by these arguments. Maybe differences in initial talent can even out over time; but what about differences in external circumstances governing opportunities for success? If these also even out over time, what about innate differences in motivation to take advantage of opportunities? Dennett might reply that one can exercise "the art of self-definition" (p. 81) to improve one's motivation. But isn't the very impetus to try to improve one's motivation itself a genetic endowment, beyond one's control? This is, of course, a reappearance of the hard determinist argument. What emerges from this formulation of it is that skill is being ruled out a priori as a cause of actions. Any achievement is chalked up to fortunate genetic endowments, or else the absence of genetic handicaps. So in answer to my earlier question, it seems that the hard determinist may indeed hold the paradoxical view of luck criticized by Dennett.

> . . . anyone who thinks that all losses are explicable in the end as due to bad luck, and all victories as due to good luck, simply misuses the concept of luck (p. 97).

The unstated premise of Dennett's argument is that the concept of luck makes sense only if it can be contrasted with instances where luck is not present. Actions may be attributed to luck only if there exist at least some other actions which are attributable to skill. In this regard, "luck" may be what Austin called a "trouser word,"[1] which makes sense only in contrast with its opposite. Dennett's argument is reminiscent of a familiar criticism of psychological egoism, the view that all actions are selfishly motivated. Its proponents explain any apparently altruistically-motivated actions by positing hidden, even sub-conscious, selfish motives. The criticism is that unselfish actions are ruled out a priori, hence rendering the concept of selfishness meaningless, in the absence of anything with which it contrasts.

Dennett aims at freeing up some elbow room for the role of skill and ability in our self-determination. As long as this elbow room exists, the existence of luck does not negate moral responsibility. His argument for restricting the role of luck can also be applied to the alleged importance of luck to the breaks and setbacks we encounter throughout our lives. As skilled self-determiners we know how to minimize the risk of being hurt by unlucky circumstances. Wearing a seat belt, and not riding on the New York City subway alone late at night are sensible

---

[1] J.L. Austin, Sense and Sensibilia (Oxford University Press, 1962) pp. 70-71.

policies. It can scarcely be called good luck if we avoid being hurt in car crashes, and being mugged. The same applies to avoiding committing atrocities. It is not good luck that prevents me becoming a mass murderer, but rather the many good habits and skills ingrained in me over the years (p. 99).

However, Dennett does admit extreme cases - psychopaths, for example - who have been so unlucky in the genetic and environmental lottery that we really don't hold them responsible. This creates the danger of a "slippery slope" - what he later calls "the Spectre of Creeping Exculpation" - ending up with everyone being excused. As I explained at the beginning of Chapter I, I consider such slippery slope arguments to be the starting point of the whole free will problem. Dennett proposes to postpone this problem until the final chapters of his book. In Chapter IV, I will argue that Dennett never convincingly deals with this slippery slope, and that his treatment of moral luck is a major weak point of his book.

d. Effective Deliberation (Chapter 5)

Allied to the concern that determinism rules out rationality, control, and self-determination is the fear that it renders our deliberations powerless. This is a major concern of Malcolm and Trusted (see Chapter I) and is captured by Dennett's image of "body English" (Section 2, above).

Since all events, from the beginning of time to eternity, have already been determined (and could be predicted by a "Laplacean demon"), it seems that none of our decisions can make any difference. The simplest (and most easily refuted) form of this fear is fatalism.[1] Describing this concern, Dennett says:

> When one learns that one's deliberation is bound to be detached in this way [i.e., 'a gear that turns without engaging any other gear'] from the critical causal paths, one does indeed discover ground for despair, apathy, or resignation (p. 105).

However the truth of determinism gives no reason to conjure up such a fearful image. Rather, Dennett was describing situations in which local fatalism operates, where our decisions can indeed have no effect. Once a suicide attempter has jumped from the bridge, his sudden decision that, after all, he wants to live is impotent. Such "pockets of local fatalism" are conspicuous precisely because of the absence of the very real causal effect that our deliberations usually have. Fatalism is the view that, since all events are determined, none of our deliberations can have any effect on the pre-ordained course of events.

---

[1] There are well known arguments that fatalism is based on a simple fallacy. For example, suppose that all events are caused. Since our deliberation is one of those very causal factors which determine our actions, how can we possibly claim that our deliberation is impotent or useless? See, for example, A.J. Ayer's "Freedom and Necessity," reprinted in Watson, op. cit., esp. pp. 22-23. Dennett's argument in this section is best viewed as a diagnosis of common fatalist errors.

Since our deliberations plainly do often influence our actions, determinism does not give rise to fatalism. Sometimes our deliberations are powerless to affect our actions, and this is when local fatalism operates. However, we need not fear that determinism implies local fatalism since, as just stated, our decisions usually do have a causal impact on our actions. Once again, incompatibilists have made an unreasonable "absolutist" demand - absence of causation - when all that is needed for freedom is absence of local fatalism.

Maybe, though, the fear can re-stated without reference to fatalism. Given that our decisions and their outcomes are already determined, isn't "soldiering on with one's deliberation" (p. 107) somehow irrational? Isn't it just an illusion that we can change the world? This fear is less well-focused, but Dennett proposes to allay it by arguing that all the conditions necessary for a "perfect deliberator" are quite possible in a deterministic world. In an extension of his naturalistic account of human action, he includes effective deliberation as a further asset which helps us to deal efficiently with our environment. In dealing with the environment, the deliberator has to divide it up into features that are fixed and changing, controllable and uncontrollable, and predictable and unpredictable.[1] Crucially, there are two

---

[1] See Elbow Room, pp. 109-12 for details.

categories of events that will always be unpredictable by us: 1) quantum-random micro-events, or else deterministic events whose causation is too complex for us to "track" them; and 2) our own future decisions.[1] Even if the truth of determinism entails that all causal paths in the future are closed except for the one that actually happens, still there are some cases in which we can never know in advance which path it will be. Hence deliberation will always have a point, since as far as we can ever tell, there are genuine options open.

> It is this epistemic openness, this possibility-for-all-one-knows, that provides the elbow room needed for deliberation (p. 113).

This view we have of the world, as containing many open possibilities for the future which we cannot know in advance, is called by Dennett our manifest image (following Sellars) (p. 111).

However, in addition to this we have a scientific image, which gives us a "more fine-grained level of description" (p. 114) of the world. This view shows us that complex events which defy explanation by our manifest image are in fact predictable, at least in principle.

---

[1] It has often been pointed out that, if we could know our future "decisions" in advance, they would not be decisions, since our mind would already be decided. By definition, a decision is the closing up of options that, until the decision, were still open. See Carl Ginet, "Can the Will Be Caused," Phil. Review, 1962; and Alvin Goldman, "Actions, Predictions, and Books of Life," American Philosophical Quarterly, July 1968, for an interesting objection.

Crucially, our "capacity to engage in real-time deliberation" (p. 114) depends on the epistemic openness provided by our manifest image. Doesn't this create the suspicion that we only manage to deliberate by consciously blinding ourselves to the objective truth given by the scientific image? It may be that we will be better deliberators by adopting our manifest image, but can it be rational to base our actions on an illusion? Dennett's response is to argue that it is not an illusion. Even in the deterministic world portrayed by our scientific image, we have "real opportunities" to influence events, and to avoid what we don't want to happen.

According to Dennett, all that is required for the ordinary concept of a real opportunity is the absence of physical obstruction and of ignorance. In contrast to the person who has just jumped from the bridge, and is subject to a pocket of local fatalism, the deliberator with a real opportunity is in a situation

> in which the outcome of its subsequent 'deliberation' will be a decisive (as we say) factor . . . the critical nexus passes through its deliberation (p. 118).

Similarly, an event is avoidable or preventable provided that the deliberator's decisions have the relevant causal effect. The reason that the death of the person who jumped from the bridge is inevitable is precisely that his deliberations are now useless: again due to local fatalism. However, despite Dennett's reassurances,

suspicions remain about real opportunities and avoidability.

There is a residual doubt that the existence of real opportunities is undercut by the truth of determinism. Dennett considers a robot, whose actions are clearly determined (pp. 118-20). If it comes up with a wrong answer, surely it never had a real opportunity to get it right, since it was merely following its pre-determined "heuristic procedure." Suppose that this procedure occasionally involves a "pseudo-random number generator,"[1] which is not quantum-indeterministic, but has the same practical effect, since it is "virtually patternless." Had this generator by chance come up with a different number, the robot would have got the right answer. However, since the generator's program determined the number, the robot never had a chance to succeed. Similarly, if determinism is true, our brains contain the equivalent of pseudo-random number generators, and we have no real opportunities when we make mistakes. Dennett fights this intuition by asking what improvement in the robot's "existential predicament" (p. 119) would be made by installing a genuinely indeterministic randomizer? Indeterminism is supposed to be more hospitable to freedom,

---

[1]For the value of random procedures in problem solving, see the section called "The Uses of Disorder," Elbow Room, pp. 66-69.

yet it would still be determined that a number will be selected. A pseudo-random generator is in practice just as impossible to predict as a genuinely indeterministic randomizer. Does the fact that the choice of an indeterministic randomizer is not already determined make it any more capable of avoiding errors? Its heuristic program will certainly be no more efficient than that of the pseudo-random generator, which is quite patternless enough to achieve its goal of avoiding a coincidental matching with the phenomena being studied.[1]

He likens the quantum indeterministic randomizer to a regular lottery and the pseudo-randomizer to a lottery in which the drawing is secretly made before the tickets are sold. In both cases luck plays the same role, and one stands the same chance of winning both. Similarly, we may view our lives in a deterministic world as lotteries in which our endowments are written on tickets which were all "drawn at once, eons ago, put in an envelope for us, and doled out as we needed them through life" (p. 121). Those who fear determinism want an indeterministic world; but the only difference this would produce would be periodic drawings, each one based on quantum-indeterministic randomness. How would this provide us with any greater chances or opportunities in life?

---

[1]See Elbow Room, p. 68, for a detailed explanation of the need to avoid such coincidental matchings between a problem-solver's program and the events it is designed to study.

The fear that no future events are avoidable is really a return to fatalism. Trivially, of course, if a "future event" can be avoided, then it is not, after all, a future event. In this harmless way, no future event can be avoided. Yet surely some catastrophes and other events are averted by human intervention, and we need to make sense of this everyday belief. In practice, we speak of an event being prevented or averted if it would have happened had it not been for a human or natural intervention. However, this is not enough. I would have died if I hadn't drunk any liquids last week, yet I would scarcely say that I thus prevented my death. We only use these verbs of "making a difference" if the event "prevented" was expected to the best of our knowledge; which my death from thirst clearly was not. Thus prevention and avoidance are in part epistemic concepts, and for them to be applicable, one must

> make a salient ('pivotal') contribution to the actual trajectory of the world - as judged by contrast with the projected or anticipated (or retrospectively judged) likely trajectory of the world (p. 126).

Clearly there are many cases of such acts of averting and preventing, even in a deterministic world. Dennett's strategy has been to show that it is local fatalism, and not determinism, that restricts real opportunities and the ability to avoid events. "Local fatalism" draws its very meaning from the fact that our deliberations usually do have a causal effect on our actions.

4. "Could Have Done Otherwise" (Chapter 6)

Up until now, Dennett's approach has been to undercut incompatibilism by means of his naturalistic account of human agency. One by one, he showed that the necessary conditions for responsibility are compatible with determinism, despite fears to the contrary. Now he directly attacks what is often viewed as the main incompatibilist argument.

(1) In order for us to be morally responsible, it must be true that we could have done otherwise than we actually did. (This is the principle of alternate possibilities, discussed in Chapter I, Section 4C.)

(2) If determinism is true, then it is never the case that we could have done otherwise.

(3) So if determinism is true, we are never morally responsible.[1]

Traditional compatibilists take the line of analyzing "could have done otherwise" in such a way that (1) is reasonable, but (2) comes out clearly false. Applying such a hypothetical analysis of this clause, (2) becomes

---

[1] I think that all incompatibilists would endorse this argument, in one form or another. The most thorough, yet also the most concise, statement of it is Peter Van Inwagen, "The Incompatibility of Free Will and Determinism," reprinted in Watson, op. cit. See Chapter I for a discussion of Van Inwagen.

(2') If determinism is true, then it is never the case that we could have done otherwise, if we had so wanted or chosen.

This is evidently false, thus blocking the unwanted inference to (3). Dennett, on the other hand, attacks both (1) and (2). First, he argues that regardless of how the crucial clause is analyzed, it has no bearing on the issue of moral responsibility, and so (1) is false. Second, this time using the notion of epistemic openness developed in the previous chapter, he argues that in a deterministic world our actions and other events could indeed have turned out otherwise than they actually did. This attacks premise (2).

Dennett's first argument may be viewed as an extended reply to Van Inwagen's incompatibilist argument. Dennett rejects the principle of alternate possibilities (PAP), which Van Inwagen assumed as the starting point of his argument. Van Inwagen argued that, if determinism is true, it is never the case that we could have done otherwise than we actually do. (This is premise (2) of my version of the incompatibilist argument.) He assumed that everyone would accept the PAP (my premise (1)), and that hence his incompatibilist conclusion was established. However, the PAP is precisely what Dennett is here attacking.

Dennett is following Harry Frankfurt, who made the first concerted attack on (1).[1] Whereas Frankfurt based his attack on a few ingenious but unlikely examples, Dennett argues that even in everyday cases moral responsibility does not hinge on whether we could have done otherwise. He gives plausible cases in which he would be unable to do otherwise, and yet would still claim to be morally responsible. For example, "it would be impossible to induce me to torture an innocent person by offering me a thousand dollars" (p. 133). (Assuming that the thousand dollars is the only incentive.) Similarly, there are many other cases in which any person with a modicum of moral awareness and self-control would have only one feasible alternative.

> Doesn't a considerable part of being a responsible person consist in making oneself unable to do the things one would be blamed for doing if one did them? (p. 135)

Furthermore, Dennett argues that it would be impossible in practice ever to know whether anyone could have done otherwise. For this purpose, he takes on the "categorical" interpretation of "could have done otherwise" which is insisted on by incompatibilists. According to this interpretation, (1) becomes

(1') In order for us to be morally responsible, it must be true that we could have done otherwise in exactly the same circumstances.

---

[1] "Alternate Possibilities and Moral Responsibility," J. Phil 65, (1969).

Thus he avoids the criticism levelled at earlier compatibilists who proferred a hypothetical analysis on the lines of (2'); namely that such an analysis dodges the main moral issue.[1] Assuming (2) to be true, whether or not a person could have done otherwise in exactly the same circumstances hinges on whether our actions are determined, or "indeterminism reigns in our brains at the subatomic quantum mechanical level" (p. 136). This issue is practically impossible to decide, since it would be incredibly hard to distinguish between a genuinely indeterministic and a Laplacean-determined effect at the microscopic level in the brain (pp. 135-36). So "the odds are very heavy that we will never have any reason to believe of any particular act that it was or was not responsible" (p. 136).

Moreover, even if we could answer the "could have done otherwise" question, Dennett claims that it would be of no moral interest. Dennett asks us to suppose that we discovered that a person's action was indeed determined, and that in exactly the same circumstances he will always do the same thing. Even if we could reproduce exactly the

---

[1]Thus C.A. Campbell, in "Is 'Free Will' a Pseudo-Problem?" argues that the hypothetical analysis gives only an "unreflective" account of responsibility, and merely postpones the challenge posed by determinism. "Could have done otherwise, if one had chosen" is all very well; but what if one's very ability to choose otherwise was determined (one way or the other)? See Chapter I, section 4a(i) for a detailed account.

same external conditions, the agent himself will never be in exactly the same state.

> People learn, and remember, and get bored, and shift their attention, and change their interests so incessantly, that it is as good as infinitely unlikely that any person is ever in the same (gross) psychological or cognitive state on two occasions (p. 137).

The information that he could not have done otherwise in exactly the same circumstances will thus be of no practical use in predicting or influencing his future behavior, since exactly the same circumstances will never again arise. Similarly, the discovery that his behavior was not determined, and hence that, in some sense, he could have done otherwise, would give us no basis for predictions of his future actions. The fact that his behavior was indeterministic does not make him more flexible or amenable to improvement.[1] In either case, the "could have done otherwise" question is "a particularly pure metaphysical curiosity" (p. 138), of no practical value.

As may be emerging from these last comments, Dennett believes that the "could have done otherwise" question is only of value insofar as it concerns our ability to judge the agent's character, and affect his future actions. This is a precursor of the openly utilitarian justification of

---

[1] See pp. 115-21 for a comparison of a deterministic and an indeterministic robot. Dennett there argues in detail that being genuinely random does not give the latter any of the extra "freedom" that we might consider necessary in humans for moral responsibility.

praise, blame and punishment which appears in the final chapter. Dennett again considers a robot with a pseudo-random number generator, which makes a mistake. Its designers know that in precisely the same circumstances, it will make exactly the same mistake. An exact repetition of these circumstances is highly unlikely, and what they are really concerned about is whether, in similar situations which may arise, the robot will make a similar mistake. If the answer is yes, they may well want to redesign its program. If the answer is no, they may be content to dismiss the error as a "don't care": even a good heuristic program involves taking short-cuts, and this creates the permanent possibility of error. "There is a difference between being optimally designed and being infallible" (p. 141). Their concern is with whether the error reveals "a pattern of systematic weakness, likely to recur" (p. 141), or was due to an unlucky coincidence.

How does this apply to human misdeeds? If we are determined, then given exactly the same micro-circumstances, we will do the same thing again. What is more important is whether our error reveals a character flaw that is likely to result in further misdeeds. If so, we try to "redesign" the malefactor, usually by means of criticism or punishment (p. 142). It is always possible to view a misdeed from "a more fine-grained standpoint" (p. 143), and to regard it as an inevitable and excusable

consequence of our state ("design") in those exact circumstances. However, especially when our own misdeeds are in question, it is much healthier to try to change our character ("redesign ourselves"), rather than to regard it as a given. We can take responsibility for our actions and character, by exercising our capacity for self-determination, described in Section B3c above.[1]

The other branch of Dennett's argument on "could have done otherwise" is his attack on an alleged consequence of premise (2) of the incompatibilist argument. This can also be viewed as an attack, by modus tollens, on premise (2). If everything is determined, then what actually happens is in fact the only possible turn of events. Dennett calls this fear, that the existence of more than one possible future is just an illusion, actualism: only the actual is possible.[2] What we need is "some [elbow] room between the actual and the possible" (p. 144). In Chapter 5, Dennett argued that all that is required for deliberation to be worthwhile, for us to have real opportunities, and for us to be capable of averting unwanted events, is epistemic openness. This means that, so far as we know, several different turns of event are possible. He now uses

---

[1] I examine both Frankfurt's and Dennett's arguments against the PAP in greater detail in Chapter III, section 5.

[2] Dennett takes this term from Ayers. See Elbow Room, p. 145, for reference.

the same concept of epistemic openness to show that the actual future is not the only possible one; and thus to find "elbow room" between the actual and the possible.

The notion of possibility which Dennett defends against actualism is epistemic possibility (p. 148). This can be illustrated by the concept of a coincidence, or accident (p. 149). It will be useful to consider an example, not given by Dennett, but taken from H.L.A. Hart and A.M. Honore.[1] Suppose that someone is punched, and falls on his back, and as he is lying on the ground a rotten tree falls and crushes his head. Provided that the tree was not caused to fall by his impact on the ground, this is a good candidate for a coincidence.[2] Now for a Laplacean demon, who would have calculated that the tree would fall at the exact moment when the victim was lying near it, his skull being crushed would have come as no surprise. However, we would still call it a coincidence, since for us it was unexpected. What counts as a coincidence is relative to our knowledge, and it is therefore an irreducibly epistemic concept. Importantly, a coincidence does not require

---

[1] Causation in the Law (Oxford, 1959), sections from which are reprinted in Feinberg and Gross (ed.) Philosophy of Law, 2nd edition (Wadsworth, 1980). This example appears in their discussion of coincidence, pp. 411-13 in Feinberg and Gross.

[2] Hart and Honore conclude that the person who punched him did not cause his head to be crushed. In their analysis, a coincidence "blocks the tracing of consequences."

absence of causation, or "absolute coincidence," as Dennett calls it. Dennett suggests that in this context possibility, like coincidence, is an epistemic concept. Provided that, to the best of our knowledge, more than one course of future events is open (i.e., coincidences and accidents may happen), then several different futures are indeed open, and actualism is defeated. This holds even if only one course of events is physically possible (p. 148), and even if a Laplacean demon would know this.

What of the objection that Dennett's notion of epistemic possibility is merely an ad hoc solution, which tries to gloss over the fact that only one real possibility always exists? He replies that for any finite being there will always be unpredictable events (see previous section) and so there will always be coincidences and possibilities of the kind he describes. Why would it be rational to abandon or restrict these concepts on the ground that they would have no application for a Laplacean demon?

> It is not any parochial fact about our epistemic limitations that distinguishes the world into stable, predictable systems and unstable, chaotic systems; it is a fact about the world itself - because it is a fact about the world's predictability by any predicting system at all, however powerful. There is no higher perspective (unless we count the perspective of an infinite being) from which the 'accidental' collisions of locally predictable trajectories are themselves predictable and hence 'no accident' after all (p. 152).

5. The Justification of our Concept of Responsibility (Chapter 7)

Having devoted most of the book to defending freedom and responsibility against incompatibilist attacks, Dennett finally turns to more fundamental questions concerning freedom and responsibility, and their importance to us. In what circumstances is it rational to believe that we are responsible beings? Is our practice of holding people responsible ever justified? And is the distinction between responsible and non-responsible people well-founded?

Nihilists, sometimes inspired by Nietzsche, call into question the justifiability of our whole apparatus of responsibility and punishment. The very compatibilist/incompatibilist debate is fruitless if the concept of responsibility is empty. To this Dennett replies that there is a purely instrumental justification of our institution of responsibility which does not depend on any metaphysical beliefs (p.155). Unfortunately, the nihilist has a more fundamental objection: nothing matters, and any activity is pointless, including the ones allegedly aided by our practice of punishment. However, this assertion involves the nihilist in a strange pragmatic paradox.[1]

---

[1] R.M. Hare gives a very interesting and graphic account of the pragmatic paradox involved in asserting the truth of nihilism in "Nothing Matters," from Applications of Moral Philosophy (Macmillan, 1972), reprinted in E.D. Klemke (ed.), The Meaning of Life, (Oxford, 1981).

If nothing matters, then there is no point in pointing out this very fact. If nihilism is true, then nothing matters, and we are free to do what we want, <u>including</u> acting as if nihilism is false. If it is false, then some things <u>do</u> matter, including arguing that it is false (p. 156).

Dennett's argument throughout the book has aimed at establishing the possibility of moral responsibility in a deterministic world. He now faces the challenge of the "slippery slope" argument which I stated at the beginning of Chapter I as the root of the free will problem. How can we justify the dividing line we draw between "exculpating pathology" and "varieties of falling short that still leave agents genuinely culpable" (p. 157)? His reply to "the Spectre of Creeping Exculpation" is a surprisingly familiar utilitarian, instrumental justification of the institution of punishment. Punishment is justified as a means of keeping social order, with penalties being "fine tuned" to maximize deterrence while minimizing the cost of enforcement. Just as Schlick and other traditional compatibilists do (see Chapter I), Dennett links responsibility to the effectiveness of punishment. To be responsible is to be amenable to deterrence - and to reform, for those unfortunate enough to commit a crime and get punished. Non-responsible people, on the other hand, are precisely those who are incapable of being influenced by our sanctions, either before or after the misdeed.

Notoriously, insane people fall into this latter group. Dennett extends exactly the same utilitarian justification to moral responsibility itself, when viewed independently of legal sanctions.[1] The only reason we blame people, including ourselves, for misdeeds is to try to prevent a repetition of similar sins.

> But whatever responsibility is, considered as a metaphysical state, unless we can tie it to some recognizable social desideratum, it will have no rational claim on our esteem (p. 163).

This passage rules out a retributive justification of punishment and responsibility, unless we make the dubious move of regarding "the restoration of justice," or some such retributive phrase, as a mere social goal. On other occasions Dennett makes statements which further alienate him from the retributivist camp. He admits that exactly where we draw the line above which people will be held responsible and amenable to punishment, and below which they will be excused, is arbitrary (p. 162). His reason is that if on any occasion we "look too closely at the particular micro-details of the accused's circumstances" (p. 162), we will tend to excuse her. Instead we want to

---

[1] Joel Feinberg also argues that it is a mistake to view moral responsibility as being the "internal" analogue of legal responsibility. ("On Being 'Morally Speaking a Murderer,'" Doing and Deserving (Princeton, 1970). Moral responsibility is not a question of "pure" guilt. It is just as dependent on external circumstances and exigencies as is the question of liability. However, Feinberg would not agree with Dennett's purely utilitarian conception of responsibility.

encourage people to take responsibility, and to take steps to improve themselves (see B3c, above). In opposition, I will argue in Chapter III that are good, non-arbitrary, retributive reasons for drawing the line governing the insanity defense and other excuses exactly where we do. Finally, he equates wrongdoing with errors of judgment. Punishment is viewed as a tax on the calculated risk taken by the wrongdoer, who hoped to gain the benefit of lawbreaking without suffering the infliction of a penalty (p. 165). This is the very concept of punishment criticized from a retributive viewpoint by Feinberg.[1] In Chapter V I will criticize Dennett's conception of punishment and responsibility, and argue that the rest of his book is compatible with a retributivist view.

Finally, Dennett returns to one of the "bugbears," "the Dread Secret." This is the fear that, if we believed in determinism, we would fear that our deliberation was powerless, and would sink into inaction. His whole book has been devoted to showing that, in spite of determinism, our decisions do control our destinies; and he has just given an instrumental justification of the social institution of punishment. However, what if the sceptic proposes that, "in the private arena of our own hearts" (p. 166), we should remain sceptical about our freedom? The social justification of punishment does not apply to

---

[1] "The Expressive Function of Punishment," Ibid.

our own private thoughts. Dennett replies that, even if one is unconvinced by his earlier arguments, there are good purely pragmatic reasons for believing that we have free will and are morally responsible. If we become preoccupied with incompatibilist fears that we do not control our destiny, our powers of rational deliberation may be impaired. Apathy may take the place of flexible, appropriate responses to our environment.[1] Since for Dennett free will consists in precisely such responses, it is correct to say that "believing that one has free will is itself one of the necessary conditions for having free will" (p. 168). Even if such freedom turns out to be an illusion, it is a "life-enabling or life-enhancing" one, and is far preferable to pessimistic apathy.

However, this pragmatic argument is incidental, and introduced only to counter fear of the "Dread Secret." The naturalistic account of human rationality and agency to which most of Elbow Room is devoted has a much more ambitious conclusion:

> What we want when we want free will is the power to decide our courses of action, and to decide them wisely, in the light of our expectations and desires. We want to be in control of ourselves, and not under the control of others. We want to be agents, capable of initiating, and taking responsibility for, projects and deeds. All this is ours, I have tried to show, as a natural product of our biological endowment, extended and enhanced by our initiation into society (p. 169).

---

[1]This is where Dennett's metaphor of "follow through" is appropriate. See my discussion of the "bugbears."

6. Reflections

As I said at the beginning of this chapter, I accept Dennett's compatibilism as fundamentally correct. Like other compatibilists, he claims to have done justice to the abilities we need for moral responsibility - e.g., rationality, control, self-determination, and effective deliberation - while assuming the truth of determinism. The usual incompatibilist objection to such analyses is that they miss out on precisely those elements of these concepts which are essential for moral responsibility. Compatibilists are accused of giving impoverished versions of these key concepts. I think that Dennett's very mode of presentation helps to forestall such criticisms. He gives very rich accounts of these concepts in a deterministic world, and persistently asks: what more could be wanted for moral responsibility?

Dennett's work successfully deflects the main thrusts of the three groups of incompatibilist objections stated in Chapter I, section 4. (1) Libertarians (e.g., Campbell) and hard determinists (e.g., Hospers, Edwards) alike argue that the fact that our desires are themselves caused undercuts our control over our actions and character. A major theme of the naturalistic view of human nature which runs throughout Elbow Room, but especially in Chapters III and IV, is that perfectly deterministic, evolutionary

processes enhance, rather than impede, our control and self-control. The natural process of evolution, and the process of maturation within an individual's life, enables him to become a better self-controller. (2) Malcolm and Trusted fear that determinism would threaten our status as agents. Our actions would be the result of physical reactions, and not the product of our rational deliberation. Dennett effectively answers this objection in Chapters II and V, where he argues that neither our rationality, nor the effectiveness of our deliberations, are impugned by determinism. (3) Van Inwagen, expressing the fear which perhaps underlies all incompatibilist arguments, argues that determinism rules out the ability to act otherwise than we do, and hence excludes moral responsibility. Dennett's reply (Chapter VI) is twofold. First, he rejects the principle of alternate possibilities which Van Inwagen assumes to be uncontested. Second, he shows that there is a sense ("epistemic possibility") in which we can indeed act otherwise than we do in a deterministic world.

However, I disagree with Dennett in two major areas, which will be the topic of the rest of this dissertation. First, his treatment of moral luck is unsatisfactory. His attempt to avoid the problem it poses, by stressing the role our ability can play in determining our lives, offers an excellent basis for a solution. However, the weakness

of his account arises in connection with the slippery slope argument: if some people are excused because of bad moral luck, why not everyone? Dennett's reply is to tie in our attributions of responsibility with a utilitarian justification of punishment. The suspicion is created that the attributions of responsibility thus based are arbitrary, and may be unfair.

Second, I argue that, even apart from its inadequacy in dealing with moral luck, Dennett's utilitarian approach is undesirable. I offer a general critique of utilitarian theories of punishment, from a retributive viewpoint. Moreover, Dennett and other compatibilists are mistaken if they believe that they have to adopt a utilitarian justification of punishment. Being a compatibilist is, on the contrary, consistent with holding a retributive view. This will be the topic of Chapter V.

I will deal with the problem of moral luck in Chapter IV. I will offer a comprehensive solution to the problem, which uses Dennett's insight, while avoiding the inadequacies caused by his utilitarian approach. But first I need to fill in a gap left in Dennett's version of compatibilism: he never proposes a sufficient condition for responsibility, which enables us to determine whether a particular person should be blamed or excused. Chapter III will be devoted to developing just such a criterion. When I consider objections to this criterion at the end of the chapter, the crucial importance of moral luck will emerge.

CHAPTER III. NECESSARY AND SUFFICIENT CONDITIONS FOR RESPONSIBILITY

1. The Slippery Slope Revisited

At the end of the first section of Chapter I, I stated two conditions which must be satisfied by a compatibilist criterion of responsibility. It must

(1) assume the truth of determinism, and

(2) count some actions as blameworthy, and some as excusable.

By definition, all compatibilist accounts meet the first condition, but what about the second? Both the traditional compatibilist view and Dennett's account defend the possibility of moral responsibility in a deterministic world. For the sake of argument, I will assume for most of this chapter that their arguments do indeed establish the general possibility of moral responsibility. However, what remains to be provided is a criterion to decide exactly when people are morally responsible. Even Dennett, whose views this dissertation is largely devoted to defending and improving, gives no precise criterion.

The crucial premise in the slippery slope argument against moral responsibility was the second one. It

asserts that there is no morally significant difference between so-called criminal, and so-called excusable actions, given that both kinds are subject to causation. This chapter will be devoted to fleshing out a positive criterion to distinguish between blameworthy and excusable actions, and hence to block the slippery slope argument.

First, I will consider an argument which occurs in Richard Taylor's article, "The Meaning of Human Existence."[1] This article is not about the determinism/moral responsibility debate, but it is a very clear example of the view of human motivation which gives rise to the incompatibilist, slippery slope argument. Taylor begins with descriptions of primitive animals whose actions are very little more than reflex responses to biological urges. For example, the mole is "programmed" to spend its days in the endless search for worms and insects, for which it has to constantly burrow underground.[2] Whatever freedom is enjoyed by the mole is far too restricted to enable us to hold it morally responsible for its actions. Similarly, Taylor describes a sensualist who, either because of being injected with a special brain fluid

---

[1] In Burton Leiser (ed.), Values in Conflict: Life, Liberty, and the Rule of Law (Macmillan, 1981). Taylor gives a similar argument in his book Good and Evil (Macmillan, 1970), the relevant section of which is reprinted in E.D. Klemke (ed.), The Meaning of Life, (Oxford, 1981).

[2] Ibid., pp. 9-10.

by the gods, or else a natural lack of self-control, is unable to refrain from his constant pursuit of physical pleasure.[1] Taylor then suggests that all human behavior is equally biologically determined. If this is part of a general determinist thesis about human behavior, then, as Dennett's arguments in the previous chapter were designed to show, it need not lead to any troubling conclusions about freedom and responsibility. The only criticism to make of Taylor in this regard is that the particular kind of determinism he advocates is a rather simplistic, Freudian view. He asserts that all of the goals of human life - e.g., "careers, the pursuit of glory, office-seeking, honors" - are ultimately just "other outlets" for our sexual urges.[2] However, what is objectionable is Taylor's conclusion that, in our everyday actions, these biological urges exert an "inner compulsion"[3] on the way we behave. This conclusion does of course threaten our moral responsibility.

Certainly the sensualist, and very likely also the mole, are correctly described as being compelled by their desires, and hence partly or completely excused from responsibility. However, even if determinism is true, there seems to be a crucial difference between these cases

---

[1] Ibid., p. 23. [2] Ibid., p. 13

[3] From Good and Evil, in Klemke (ed.) op.cit., p. 148.

and our everyday actions. His picture of us as "prisoners of our own desires," unable to escape them, makes a crucial mistake: it identifies motivation with compulsion. Doing those things that we want to do seems to be the clearest possible case of lack of compulsion. Surely the concept of compulsion only makes sense if it can be contrasted with examples of freely chosen, purposive actions. Taylor's view of motivation seems to rule out the possibility of the latter a priori, and hence to render "compulsion" meaningless.[1] The preceding criticism of Taylor is parallel to the familiar compatibilist doctrine that causation needs to be distinguished from compulsion. (See Chapter I, Section 3a.) Just as incompatibilists wrongly believe that causation implies compulsion, Taylor wrongly holds that compulsion is implied by motivation.

2. Desires and Compulsion

I think that compatibilist replies of the kind just given to Taylor-type incompatibilist arguments are conclusive. However, the mere assertion that not all of our motivations are compulsive does not yet provide a criterion for determining which motivations are and which are not compulsive. The

---

[1] See Joel Feinberg, Doing and Deserving (Princeton, 1970), p. 290, for a similar application of this argument.

virtue of Taylor's argument is that it draws attention to the need for such a criterion.

Dennett himself provides a useful first approximation of a distinction between responsible and excusable actions.[1] He argues, as we saw, that the "benign" causation of our actions - i.e., the adducement of information and reasons relevant to our actions - in no way rules out responsibility. In these cases, the causation operates via our rationality, and hence leaves us in control of our actions. In contrast, the type of causation which does indeed excuse us from reponsibility - e.g., hypnosis, deception, coercion - does so precisely because our rationality is by-passed. We have no opportunity to act on our own beliefs and desires. Dennett's distinction is indeed useful in distinguishing free, responsible actions from the most blatant forms of compulsion. However, it is incomplete, in that it does not take satisfactory account of what we may call "compulsive desires."

Dennett's account is vulnerable to a criticism often levelled at traditional compatibilist accounts. According to Schlick, our actions are free provided that we do what we want to do, without undue influence from other people. (See Chapter I, Section 3b.) Doesn't it follow from this

---

[1] "Mechanism and Responsibility," Brainstorms. See above, Chapter II, section A3 for a detailed account of Dennett's analysis.

that kleptomaniacs, pyromaniacs, the sensualist described in the previous section, and other people subject to compulsions, will also count as free, responsible agents? After all, they are all doing what they want to do.[1] Similarly, such people would appear to qualify as responsible, according to Dennett's criterion. Their rationality may play a key causal role in their actions: the kleptomaniac, for example, may scrupulously avoid being caught. The reason why we tend not to view such people as responsible agents is that, though they act willingly and on the basis of their own decisions, they are unable to resist the desire which motivates them, even if they wanted to. The existence of such people shows that compulsion can be exerted not only by other people, but also by our own desires. Schlick does recognize cases similar to the compulsive ones we are considering, and attempts to reconcile them with his criterion of responsibility as the uncoerced exercise of our will. For example, he says of the mentally ill person that he is not responsible "because we view the illness as a disturbing factor which hinders the normal functioning of his natural tendencies."[2] Schlick never explains how to determine whether an action

---

[1]For an example of this criticism, see Sir Isaiah Berlin, quoted by Gary Watson in "Free Agency" in his Free Will, p. 97.

[2]Moritz Schlick, in Edwards and Pap, op. cit., pp. 62-3.

is a "normal" expression of someone's "natural tendencies." To make such decisions we need a precise criterion to distinguish compulsive from noncompulsive desires.

The extension of compulsion to cover not only external influences, but also some of our desires themselves, requires a deeper analysis of freedom than the mere unhindered exercise of our desires. This need is recognized by some traditional compatibilists, for example A.J. Ayer. As we saw in Chapter I, Section 3b, he points out that the kleptomaniac does what he wants to do, without acting freely. What is needed for freedom and responsibility is the ability to refrain from acting, as well as to act on, our desires. This deeper analysis of freedom and responsibility, on which neither Ayer nor any others of the traditional compatibilists give any detail, needs to provide for our control over our desires. The reason that we tend to excuse compulsive people for their faults is that they "can't help" following their desires.

A reflective version of compatibilism, then, must require not only control of our actions, but also control of our desires. To be free from compulsion, we must be able to choose which of our desires to act upon. We must be able to resist a desire if we decide that it would be prudent to do so; and, more generally, we must be able to organize and tutor our desires, encouraging healthy ones

and phasing out destructive ones. For example, Mill gives a very plausible account of moral education as the process of acquiring good habits. When we bring up children, we use rewards and punishments in such a way that eventually they want to do the right thing. This process can be applied to ourselves as an effective method of self-control. (See Chapter I, Section 4a(ii).)

In Chapter II, Section B3c, I discussed Dennett's account of self-determination, or self-definition, as he calls it. His discussion, in Chapter 4 of Elbow Room, is of great value in dispelling fears that determinism would rob us of self-control. He also brings out the intimate connection between self-determination and responsibility by drawing on Charles Taylor's notion of "strong evaluations," which I intend to discuss in detail below. However, Dennett does not answer the question I set out at the outset of this chapter: how do we distinguish blameworthy from excusable actions? Alternatively, exactly when are we in control of our desires, and when are we compelled by them? In the light of the considerations raised in this section, the answer to this question will require a detailed analysis of desires, compulsion, and self-control. What distinguishes the self-controlled, responsible person?

3. An Analysis of Self-Determination, and a Criterion of Responsibility

Realizing the incompleteness of the traditional compatibilist account of desires and freedom, several philosophers have recently focused on the process by which we control ourselves. I will look at Harry Frankfurt's "Freedom of the Will and the Concept of a Person,"[1] Gary Watson's "Free Agency,"[2] and Charles Taylor's "Responsibility for Self."[3] I will draw from these articles a general picture of self-control, and using this picture suggest a condition which is at least necessary for moral responsibility. Anyone not capable of such self-control is not even a candidate for moral responsibility; while for those who can exert it there is, ceteris paribus, a presumption of responsibility.

Frankfurt's discussion is particularly well suited to our purposes in this section. He presents his article as an account of what it is to be a person, that is a being with

> those attributes which are the subject of our most humane concern with ourselves and the source of what we regard as most important and problematical in our lives.[4]

---

[1] Journal of Philosophy, January 1971.

[2] Journal of Philosophy, April 1975.

[3] Amelie Oksenberg Rorty (ed.), The Identities of Persons (University of California Press, 1976). All three articles are reprinted in Watson, op. cit., from which all page numbers are taken.

[4] Frankfurt, in Watson, op. cit., p. 82.

The particular attribute with which we are concerned is the capacity for moral responsibility. Only a person is capable of moral responsibility, and "both of enjoying and of lacking freedom of the will."[1]

The concept of a person is best understood in contrast to that of a wanton, that is a being who has no interest in controlling or evaluating his desires. Frankfurt distinguishes between two levels of desires. First-order desires are simple wishes, preferences, or inclinations. Second-order desires are desires about our first-order desires. For example, I may decide that my penchant for expensive meals is inconsistent with my concern for the plight of starving Ethiopians. As a result, I will restrain my desire for food, and send larger donations to famine relief organizations. Alternatively, my motive for restraining my gastronomic indulgence may be my judgment that it is unhealthy or extravagant. All of these second-order desires are desires about the desirability of our first-order desires. A wanton is capable only of first-order desires, whereas a person is capable of both kinds. More specifically, Frankfurt holds that what distinguishes a person is the capacity for second-order volitions, that is desires that a certain first-order desire be effective in motivating our actions.

---

[1]Ibid., p. 89.

The clearest examples of wantons are lower animals, and human infants. They merely follow their desires, and their inability to question or restrain their inclinations seems to be precisely why we do not hold them responsible. However, some wantons <u>are</u> capable of considerable debate concerning how best to satisfy their many, sometimes incompatible, desires. We may suppose that the sensualist discussed in Section 1 of this chapter spent much time carefully deliberating on how to maximize his sensual pleasure. He may, for example, decide to moderate his eating and drinking, so as not to interfere with the pleasure of sex. What makes him a wanton is that "he is not concerned with the desirability of his desires themselves."[1] His debate is at the level of rival first-order desires, and he makes no second-order volitions based on the superior intrinsic worth of certain desires over others. Desires are counted solely in <u>quantitative</u> terms.

Why being a person is relevant to moral responsibility is that exercising second-order volitions enables us to either endorse or disown our first-order desires. Whereas a wanton merely follows the current of his first-order desires, with a few adjustments here and there to maximize their satisfaction, a person is more truly master of his

---

[1]Ibid., p. 87.

destiny. However, the objection arises that the problem posed by determinism has only been postponed, and pushed back to the level of second-order desires. Aren't we the passive observers of the conflicts that exist between our second-order desires, and no more responsible for them than a wanton is for following his first-order desires? As Watson objects, "Can't one be a wanton, so to speak, with respect to one's second-order desires and volitions?"[1] Maybe we could make our decisions more fully our own by means of third-order volitions, but there seems nothing to prevent an infinite regress of meta-level desires. Frankfurt replies that we only need to ascend to second-order volitions to "identify ourselves decisively" with a first-order desire. Ascending to the second level is enough to assert our self-control, and renders pointless any higher-level volitions.[2]

Watson contends that Frankfurt's reply is ad hoc and question-begging. There is no reason why our desires at any meta-level should be regarded as the ultimate expression of our character, to which we are committed. If such conclusive "identification" does occur, it can only happen if weightier factors than our desires enter into our deliberation. (Watson calls these "weightier factors," as we shall see, our "values.") Though Watson thus rejects

---

[1] Ibid., p. 108.

[1] Ibid., pp. 91-2.

the worth of Frankfurt's notion of different levels of desire, he is in fact in broad agreement with Frankfurt's general goal. Both Watson and Taylor, in their accounts of freedom which I will consider next, produce distinctions which closely correspond to Frankfurt's between persons and wantons.

Rather than consider the wanton and person as two different beings, Watson picks out two different sources of motivation within a single person: one's desires, and one's values. A person's desires flow from or consist in appetites or passions "whose existence and persistence are independent of the person's judgment of the good."[1] A person's values, on the other hand, are the standards by which he judges his desires themselves, among other things. They are

> those principles and ends which he - in a cool and non-self-deceptive moment - articulates as definitive of the good, fulfilling and defensible life.[2]

Very often the things we value are also desired by us: in fact, the sign of a virtuous person may well be that he wants to act well. However, it is also possible to value things we do not desire. Thus I may well judge that I ought to visit an elderly relative in hospital, while considering it a rather stressful chore. Crucially, desiring something does not imply valuing it. What

---

[1] Ibid., p. 102. [2] Ibid., p. 105.

distinguishes desires from values is thus that we can desire something without considering it desirable.

Frankfurt's wanton is always motivated by his desires only, and has no valuation system at all. We are all wantons with respect to those of our motivations which proceed solely from our desires, in opposition to, or at least independently of, our values. It is the potential conflict between our desires and values that creates for us the problem of free well. For wantons, no such conflict is possible, and hence no problem exists. According to Watson, we are free, and hence responsible, only when we are capable of being motivated by our values. These values are not "meta-level" desires; rather they are different sources of motivation. In support of his analysis of freedom, Watson points to our subjective experience of deliberation concerning our future actions: even in our deepest self-questioning, we typically debate on how to act, and not on what we should desire.[1] For example, it seems more accurate to view my concern for starving people as a direct motivation for my giving up expensive meals. It is a rival motivation to my desire to eat expensive food, rather than a "second-order desire" about my desire to eat such food. My concern for starving people differs in that it is something I value, whereas my gastronomic inclinations are mere desires.

---

[1] Ibid., 109.

Animals and babies cannot be held responsible by Watson's criterion because they simply have no values. As for the sensualist, even though in other areas of his life he may be responsible, he appears to have no values with regard to his pursuit of pleasure. If he were unhappy about his addiction, he would at least have values about it; but his addiction would make him incapable of acting on his judgments of value, so he would still be excused. Of course, the sensualist may be blamed for having become addicted to physical pleasure in the first place. At some time in the past, we may safely assume that his values were able to motivate his behavior, but that he simply failed to exercise them, and instead merely went along with his desires.

Corresponding to Frankfurt's wanton, and Watson's being who is motivated only by his desires, Taylor describes the "simple weigher of alternatives." The latter only deliberates on how to maximize the satisfaction of his desires. Corresponding to Frankfurt's person, and Watson's notion of the being whose values motivate his actions, is Taylor's concept of the "strong evaluator."[1] The latter is not only concerned with satisfying his desires, but also engages in "qualitative reflection" on the value of his pursuits and lifestyle. Being a strong evaluator, Taylor

---

[1]Ibid., p. 116.

argues, is necessary for being a morally responsible person. The virtue of his account is that he goes further than Watson in articulating the intimate connection between strong evaluation and moral responsibility.

First, Taylor considers the possibility of viewing all of our deliberations through the eyes of a simple weigher of alternatives. He admits that such a view, which is likely to be held by a hedonistic utilitarian, is possible. Thus my decision to stop going out for expensive meals, and to send the money I save to famine relief, would be explained as a calculation that this will increase my overall welfare. I will have a clear conscience, and doubtless my health will improve through my moderation. However, Taylor suggests that such a simple weigher of alternatives would be "an impossibly shallow character."[1] His indifference to the moral and intellectual quality of the lifestyle to which his hedonistic calculations lead him ignores the deeper questions faced by human beings. The reason why I think we would criticize a person who adopted such an attitude is that he, unlike animals and infants, is capable of being a strong evaluator, but fails to use his potential. (It is interesting to note that Dennett quotes, with approval, Taylor's arguments for the importance of strong

---

[1]Ibid., p. 117.

evaluations. As will emerge later in this chapter, and in subsequent chapters, Dennett is himself a utilitarian who may well be vulnerable to Taylor's charge of shallowness.)

Second, Taylor emphasizes the fact that our strong evaluations concerning courses of action and lifestyle are our own creation. Our concern in this section, it will be remembered, is to give an analysis of self-determination or self-control which makes clear why we can be held responsible. The underlying intuition was that, although our desires may be "given," determined by factors beyond our control, we can take responsibility for ourselves by acquiring, shaping, or suppressing desires. What makes us particularly responsible for our strong evaluations is the way they contrast with straightforward descriptions, whose only goal is to accurately portray objective reality. In contrast, the goal of our strong evaluations is rather to "articulate" our sense of what is important, which may be "initially inchoate, or confused, or badly formulated."[1] Crucially, the values which we seek to articulate are of our own making, and the very act of articulating them in part constitutes our values.

> To give a certain articulation is to shape our sense of what we desire or what we hold important in a certain way.[2]

---

[1] Ibid., pp. 122-3. [2] Ibid., p. 123.

The reason for this is that what we are assessing in strong evaluations are, in part, our very criteria for making these evaluations.

To illustrate this feature of radical evaluations, consider someone who is wondering whether it would be patriotic to refuse to register for the draft. Such a person is evaluating the moral desirability of his actions, and is to that extent a strong evaluator. However, his ruminations may lead him to a deeper level of evaluation, and he may question whether patriotism is itself a desirable quality. Since this self-questioning can ascend to indefinitely high levels of meta-ethical debate, our articulations of our values "are intrinsically open to challenge."[1] What is emerging from this discussion is that there are at least two levels of strong or radical evaluation. The minimum level required to qualify as a strong evaluator is the ability to question the moral justifiability of one's actions. This much is exhibited by the person pondering the rightness of draft evasion. A deeper level of evaluation is exhibited when he questions the value of patriotism, and enters into meta-ethical debate. It is the possibility of the deeper kind of strong evaluation, in which we define our own criteria of assessment, that maximizes our power of self-determination,

---

[1] Ibid., p. 124.

and makes us morally responsible. Conversely, the failure to make or act on strong evaluations is blameworthy in any being who has the capacity for them. At the same time, it should not be thought that our evaluations are arbitrary. They have to be faithful to the judgments of worth that we already make. It is because these judgments are often vague and unformed that our strong evaluations can be said to create our values.

Despite differences in terminology and the formal apparatus they use, I think that Frankfurt, Watson and Taylor make exactly the same distinction between exactly the same concepts. I will try to capture the distinction common to all three accounts, introducing my own terminology. Frankfurt's first-order desires, Watson's desires, and Taylor's simple weighings of alternatives are all mere preferences or inclinations, which stand in no need of moral justification. These, or rather our deliberations on how to satisfy them, I call quantitative evaluations. The only question which arises in regard to them is how best to satisfy as many of them as possible. These are the motivations which guide us through everyday life when no moral questions arise. In contrast, Frankfurt's second-order desires, Watson's values and Taylor's strong evaluations are moral judgments. These I call qualitative evaluations, in that they are an assessment of the desirability of a course of action, as

opposed to the mere question of whether we desire it. These evaluations assess the intrinsic worth of our motivations, as opposed to their intensity and the degree to which they are compatible with our other motivations.

Being capable of quantitative evaluations is unquestionably a necessary condition of being morally responsible. A being without desires or preferences is scarcely capable of action, let alone moral responsibility. Equally, the capacity for qualitative evaluations is clearly necessary for responsibility. Animals and human infants are capable of desires (and possibly also of quantitative evaluations), but cannot make any moral judgments (qualitative evaluations). This is precisely why we do not hold them responsible for following their desires without any moral restraints. I want to go one stage further and assert that the capacity for acting on qualitative evaluations is both necessary and sufficient for moral responsibility.

A person is responsible for an action if and only if:

1. It is the result of a morally justifiable qualitative evaluation, or
2. He is capable of basing it on such an evaluation. This includes
    a. The person who instinctively acts morally correctly, but who could, if pressed, form a

qualitative evaluation which justifies his action, and

b. The wrongdoer who fails to act on a morally justifiable qualitative evaluation, because either

(i) He forms an unjustifiable qualitative evaluation, or even none at all, or

(ii) He forms a justifiable evaluation, but acts against it, knowing his action to be wrong.

The first condition applies to the healthy, reflective, and morally steadfast person who is perhaps rarely encountered in everyday life. Still, most of us do fit this description on occasion: namely, when we recognize a decision as a moral one, and base our actions on our moral judgment. However, this condition alone would be plainly inadequate, since it would be absurd to suggest that the mere failure to act on a justifiable evaluation excuses one from responsibility.

The second condition avoids this unwanted consequence by recognizing that the capacity for qualitative evaluations is sufficient for responsibility. In particular, part b holds the wrongdoer responsible for his culpable failure to act on the appropriate qualitative evaluations, of which he is perfectly capable. Such a person may negligently or recklessly fail to engage in the

moral deliberation which would show him the wrongness of his proposed action (part b, section (i)). Drunk drivers, and other rather impulsive wrongdoers would be examples of this. Alternatively, the wrongdoer may be fully aware of the wrongness of his proposed action, yet callously proceed with it anyway (part b, section (ii)). This brings to mind the hardened criminal, who has learned to ignore the voice of his conscience, and routinely does what he knows to be wrong.

In contrast, some people may be genuinely incapable, in a given situation, of acting on a morally justifiable qualitative evaluation. Someone in this situation will meet neither of my conditions for responsibility, hence qualifying for being excused. I will show in the next section how my criterion of responsibility in this way accounts rather well for our actual practice in excusing people.

Part a of the second condition is needed to obviate a criticism to which my criterion, along with that of Taylor, would otherwise be vulnerable. The first condition presents an overly intellectual picture of human action; whereas in reality most of our actions, even our praiseworthy ones, are preceded by <u>no</u> qualitative evaluations.

Analogous to Taylor's strong evaluations, there are two levels of my qualitative evaluations. The first, and most

common, level is moral deliberation about how to act, e.g., should I have an abortion? Would it be dishonest to "pad" my curriculum vitae with the titles of insignificant term papers I have written? At a deeper level, we may engage in meta-ethical debate about the justifiability of our values themselves. (Taylor argues that this process of creating our own values is especially important for our moral responsibility.) By this, I do not mean the analysis of ethical concepts and language done by moral philosophers. Rather, I am referring to a process of which laymen are equally capable: questioning the justifiability of one's moral beliefs. Thus the protagonists in my earlier examples may question whether all human life is intrinsically valuable, or whether honesty is always required. The criticism obviated by part a of the second condition is that we frequently engage in neither level of qualitative evaluation, yet no one doubts our moral responsibility for our actions.

Many of my everyday actions are instinctive, and are preceded by no moral deliberation at all. Thus I drive my car on the right side, within the speed limit, and pay careful attention for reckless pedestrians. What makes me responsible for my actions is that I could, if questioned, justify my actions by reference to my concern for my own and others' safety. By the same token, this very capacity to morally appraise my actions would make me blameworthy if I were to drive dangerously.

However, another part of the objection is harder to dismiss. It may be granted that nearly everyone has the capacity for moral deliberation. But does the average person ever actually engage in meta-ethical debate, and furthermore is it reasonable to expect her to do so? Isn't there a danger that regarding the capacity for meta-ethical debate (an integral part of qualitative evaluation) as necessary for responsibility will automatically exclude most people from moral praise or blame? Most people, surely, inherit their parents' moral values, or else acquire them from church, peer groups, or from the media in general. This tends to be a matter of unreflective acceptance, with little or no debate as to the worthiness of these values.

In reply, it must first be recalled that "meta-ethical debate" refers not to esoteric philosophical analysis, but rather to the more straightforward questioning of values which any rationally competent adult may be expected to understand. Still, is it reasonable to expect people to exercise this capacity which few people actually use? Fortunately, in most cases the moral values unreflectively accepted by people are singularly good ones. Almost all religious, ethnic and political creeds are united in condemning murder, rape, robbery and other misdeeds. There seems to be no injustice involved in holding responsible those who base their actions on the blind acceptance of

these moral beliefs, especially since anyone has the capacity to question these beliefs.

More difficult is the case of the person who bases his actions on the blind acceptance of morally repugnant beliefs. Do we hold responsible the Ku Klux Klan member who commits racist atrocities based on the beliefs which he has inherited from his parents? If there is clear evidence that he is mentally incompetent, or otherwise acting involuntarily, he will be excused. However, what is far more likely is that we will tell him that he jolly well ought to have questioned and rejected the repulsive beliefs his parents gave him. As a rational adult, he is capable of such meta-ethical debate, and this is a good example where we will hold someone responsible for his actions precisely because of his failure to exercise this capacity. This is the very approach to the assessment of wrongdoers' responsibility advocated in condition 2, section b(i). The fact that he acts on a mistaken moral judgment does not excuse him from responsibility if he was capable of detecting the mistake by means of meta-ethical debate (as stated at the beginning of condition 2).

I conclude that it is sometimes just to hold people responsible, even in the absence of qualitative evaluations, either in the form of moral deliberation or meta-ethical debate. What is necessary (and also sufficient) for a person to be responsible for an action is

that she be capable of basing it on a justifiable qualitative evaluation.

4. Application of the Analysis to Traditional Excuses

The force of the criterion of responsibility I have just defended is that the capacity for qualitative evaluations is necessary and sufficient for responsibility. However, it should be remembered that I also stated that the capacity for quantitative evaluations (i.e., nonmoral deliberation on how to satisfy our desires) is necessary for responsibility. It is just conceivable that some angelic, ethereal creatures could be capable of qualitative evaluations, while having no desires. For earthly beings, it is safe to assume that the capacity for qualitative evaluations presupposes the capacity for quantitative evaluations. Accordingly, moral responsibility can be negated not only by an inability to act on justifiable qualitative evaluations, but also, in a more fundamental way, by an inability to act on quantitative evaluations. Thus, for example, a plant's lack of responsibility is best explained by its lack of desires, without reference to its lack of moral judgments. What I want to show in this section is that my criterion of responsibility is consistent with, and helps to explain, our practice in excusing people.

a. Ignorance

The excuse of ignorance applies when one's deliberations are disengaged from one's actions, or their outcome. Of course, one may not deliberate at all when acting under ignorance. The crucial point is that, even if one had deliberated, one would have been unable to prevent the untoward event. This can often be explained without reference to qualitative evaluations, since even one's quantitative evaluations may be powerless. (Since these are necessary conditions of qualitative evaluations, it follows that the latter are also powerless, thus negating responsibility by my criterion.) For example, suppose that I serve a guest with a drink which, unknown to me, has been laced with poison. Her death will follow regardless of any decision I make, or can reasonably be expected to make. I can serve it as gracefully and carefully as I like, but given my state of ignorance, nothing I do can prevent her death. I am excused because none of my evaluations can influence my behavior in this regard.

An intuitive reason as to why we excuse both animals and human infants for all of their actions is that "they don't know any better." (Though, of course, especially with animals, we may use Pavlovian conditioning to try to influence their behavior.) Let us suppose for the sake of argument that they are capable of (quantitative) deliberation about how to best satisfy their desires. Even

if this is possible, it remains clear that they are incapable of qualitative evaluations, and this is precisely why we excuse them. Such beings don't understand moral concepts at all, and are destined, as it were, to follow wherever their desires and inclinations lead them.

b. Compulsion

As in the case of the ignorant person, the actions of the compelled person are disengaged from her deliberation. Once again, such a person might not deliberate at all. What matters is that, even if she were to deliberate, she could not prevent the impending harm. Also similar to ignorance, the reason why compulsion is an excuse can often be explained by the impairment of quantitative evaluations, without reference to the (lack of) capacity for qualitative evaluations.

Someone who is physically forced into a car when being kidnapped, or who has an epileptic fit, has no control over her behavior. Not even her quantitative evaluations can affect her behavior. The case of the bank clerk who is forced at gunpoint to hand over cash is slightly more complex, but can be handled similarly. Her quantitative evaluations are a causal factor in her action, since she does choose to hand over the cash rather than be killed. However, this is only because her options have been severely restricted by the coercion of the bank robber. All things considered, she would far prefer to neither hand

over the cash nor be killed. This quantitative evaluation of hers (which could also be viewed as a qualitative evaluation, since she may regard protecting the bank's money as a matter of duty) is prevented from determining her actions by the duress created by the robber. Such an analysis can be used to cover most cases of compulsion exerted by other people.

My analysis of responsibility in terms of the capacity for acting on qualitative evaluations comes into its own in cases where compulsion is allegedly exerted by a person's own mental state or desires. Someone who is drunk or under the influence of a psycho-active drug may be unable to control his actions. Nonetheless, he is doing what he wants to do, and in this way his actions are the result of his quantitative evaluations. Notoriously, he and other people who get drunk may regret their actions the next day. This is because they are now under the control of their qualitative evaluations which, while they were drunk, were ineffective. In other words, the reason why we to some extent excuse people for actions performed when they are drunk or in other ways drugged is that they are temporarily unable to act on their qualitative evaluations. However, they will of course be blamed for not acting on these evaluations before becoming incapacitated. A similar analysis can be applied in cases where people are subject to strong, uncontrollable desires. A woman who, after constant abuse and provocation

from her husband, murders him (see Chapter I, section 1), is acting on a quantitative evaluation, by carrying out her desire. However, as her subsequent remorse will very likely show, she was not acting on a duly considered qualitative evaluation, i.e., she will admit, in a cooler hour, that the murder was wrong. To the extent that we judge that she was incapacitated, by her husband's provocation, from acting on her qualitative evaluations, we will excuse her.

c. Insanity

Insanity is frequently viewed as an excuse by virtue of the two excuses already discussed. Crazy people often don't know what they're doing, or that what they're doing is wrong (ignorance), or else they are unable to control their actions (compulsion). These two traditional excuses are reflected in most formulations of the insanity defence. For example, the current rule in most Federal courts in the United States is the American Law Institute's Model Penal Code, which was adopted by the U.S. Court of Appeals in U.S. v Brawner (1972).

> A person is not responsible for criminal conduct if at the time of such conduct as a result of mental disease or defect he lacks substantial capacity either to appreciate the criminality [wrongfulness] of his conduct or to conform his conduct to the requirements of the law.[1]

---

[1] United States v Brawner (U.S. Court of Appeals, D.C. Cir., 1972).

However, I want to consider the view, which is emerging in the literature, that insanity is related to but distinct from the traditional excuses. Donald Hermann has argued convincingly that a genuinely insane person may fail to qualify for the traditional excuses of ignorance and compulsion. He may meet the requirement of mens rea (i.e., he may be in the mental state required by the definition of the crime) by voluntarily performing a proscribed act. However his action, though subject to neither ignorance nor compulsion, is the result of his mental illness, and it seems clear that such a person may still be excused, because of his insanity. Hermann suggests that we regard insanity as a separate ground for excuse, analogous to self-defense, which also operates even if the action is fully voluntary. Both excuses undercut responsibility despite the intentional, uncoerced commission of a crime. Even though an insane person may voluntarily commit a crime, his general capacity to conform his conduct has been impaired by his mental illness. Though he may have the requisite beliefs and intentions which define a specific crime, these beliefs and intentions may themselves be the blameless product of his mental illness.[1]

The view that insanity should be regarded as a separate excuse is very well defended by Michael S. Moore, who goes

---

[1]Donald Hermann, The Insanity Defense: Philosophical, Historical, and Legal Perspectives (Thomas, 1983).

so far as to assert that "madness itself precludes responsibility."[1] What leads us to call someone insane, according to Moore, is that we find his "pattern of past behavior unintelligible in some fundamental way."[2] We are able to describe the behavior of a psychologically healthy person in terms of a practical syllogism. She has certain desires, believes that a particular course of action will satisfy those desires, and acts accordingly. An insane person, on the other hand, may have desires which we find hard to understand, or which blatantly conflict with her other desires and interests. What on earth propels an exhibitionist to expose himself in public? Even if a psychologist could find the unconscious motive behind it, it is plainly an irrational desire which works against virtually any other desire he may have. Another way in which the actions of an insane person defy explanation in terms of a reasonable practical syllogism is that the actions may be hopelessly inappropriate to the goals she pursues. A recent, infamous example of this was John Hinckley's extraordinary belief that killing Ronald Reagan would impress the actress Jodie Foster. In various ways, an insane person's beliefs and desires are themselves irrational, or her actions cannot be understood as a

---

[1] "Legal Conceptions of Mental Illness," in B.A. Brody and H. Tristram Englehardt, Jr. (eds.), Mental Illness: Law and Public Policy (Reidel, 1980), p. 62.

[2] Ibid., p. 58.

rational attempt to act on these beliefs and desires.

Why does this rule out responsibility? The reason is that such people do not even qualify for membership in the moral community. Those who are excused because of ignorance or compulsion, on the other hand, remain perfectly rational and intelligible to us. In most areas of their lives they can be held responsible. Being guided by reason, argues Moore, is a pre-condition for being eligible for moral responsibility.

> Only if we can see another being as one who acts to achieve some intelligible end in the light of some rational beliefs will we understand him in the same fundamental way that we understand ourselves and our fellow men in everyday life. We regard as moral agents only those beings we can understand in this way.[1]

Joel Feinberg further elucidates why we reduce or withhold responsibility from those whose actions we do not understand. He compares the mentally ill offender with the rational, self-interested criminal whose motivations are perfectly clear, and whom we don't hesitate to blame.[2] One of the reasons we punish the latter group is that they profit at the expense of other people, taking more than their fair share of society's benefits. Punishment, according to a retributive view, aims in part at taking back from them what they have unfairly gained. In contrast, the insane person often gains nothing at all from

---

[1] Ibid., p. 61.

[2] "What Is So Special about Mental Illness?" Doing and Deserving (Princeton, 1970), pp. 280-9.

his offense, since his behavior has no coherent or realistic goal. Consider, for example, the exhibitionist or child molester whose actions cause themselves great anguish. Since the crimes of insane people often serve no interest at all, it would be gratuitous to add to the suffering which merely committing the crime very likely caused the criminal himself to endure. Furthermore, insane people may be in the dark about their own motives. The child molester may view his deviant desire as an alien impulse: he wants to do it, but cannot explain why. Blaming someone who is puzzled by his own motives, especially if he regrets his actions, seems to be unjustifiably strict.

How does my criterion of responsibility, based on the ability to act on justifiable qualitative evaluations, apply to insane people? It is likely that the most extreme case, the psychopath, is incapable of any qualitative evaluations. He follows his own self-interest, acts only on quantitative evaluations, and may well be incapable of appreciating the moral restraints we expect him to obey. Like Frankfurt's wanton, questions about the desirability of his lifestyle do not occur to him.

What about the insane person who, unlike the psychopath, is capable of forming qualitative evaluations, but whose mental illness impairs his reasoning in one of the ways described above? Take for example John Hinckley, whose rather sophisticated plans and goals suggest that he

did value some states of affairs as worthy goals (e.g., impressing Jodie Foster). We could deny that such an impulsive goal counts as a qualitative evaluation, and view him as a wanton, being led on and confused by his conflicting desires. Alternatively, we could assert that his delusions prevented him from taking realistic steps to achieve his goal. In general, the insane person genuinely suffering from the breakdown in practical reasoning described above will be excused by my criterion because either

1. He will be unable to form any justifiable qualitative evaluations, or
2. He will be unable to follow a coherent and realistic course of action in accordance with his qualitative evaluations.

Hence, by being unable to form justifiable qualitative evaluations, or unable to act on them, many insane people will lack the capacity which, according to my criterion, is necessary for responsibility.

However, Feinberg considers some puzzling cases of mentally ill criminals who will prove to present particular problems for my analysis. These are "mentally ill but rationally competent offenders"[1] whose reasoning powers are not affected in the same way as the insane people described above. Their mental illness, in contrast,

---

[1] Ibid., p. 291.

affects their control over their volitions, e.g., kleptomaniacs, arsonists, exhibitionists, etc. Consider the arsonist, perhaps a troubled adolescent, who persists in setting buildings on fire, even though he doesn't even know the inhabitants. He realizes that what he is doing is both imprudent and morally wrong, and always regrets his acts of arson. He often manages to resist the urge to set fire to houses, but not always. I think we would want to diminish his responsibility, and Moore's account in terms of practical reasoning justifies this. Even though the arsonist's reasoning (i.e., his ability to draw inferences) is intact, the very irrationality of his desire to burn buildings (i.e., the way it conflicts with other, more important interests) resists explanation in terms of a reasonable practical syllogism.

Unfortunately, it is not clear that he would be excused by my criterion. He acts voluntarily, and makes qualitative evaluations which condemn his acts of arson. Furthermore, he suffers from no delusions or breakdown in rationality which prevent him from acting on his qualitative evaluations. After all, there are times when he restrains himself from starting fires. How is he any different from the rational, self-interested criminal, such as the bank robber, who knows that what he is doing is wrong, but does it anyway? Both people, we are tempted to say, could and should have resisted the temptation to do wrong, by following their qualitative evaluations.

The obvious reply is that there is a crucial difference between the arsonist and the bank robber. The arsonist is possessed by a crazy, compulsive urge, whereas the bank robber calmly decides to run the risk of arrest in order to get rich fast. However, as Feinberg points out, such a distinction is based on a common misconception about abnormal desires.

> There is no a priori reason why the desires, impulses, and motives that lead a person to do bizarre things need necessarily be more powerful or compulsive than the desires that lead normal men to do perfectly ordinary things . . . strictly speaking, no impulse is "irresistible." (my emphasis)[1]

Whether a desire can be resisted on a particular occasion depends on the circumstances, and not on the nature of the desire. It is quite conceivable that a perfectly "normal" desire could in practice be irresistible, e.g., the desire to drink water of a man who reaches an oasis after several days in the desert. However, "for every case of giving in to a desire . . . it will be true that, if the person had tried harder, he would have resisted it successfully."[2] This applies just as well to abnormal as to normal desires, however unreasonable it might be to expect people to resist their desires in certain circumstances.

This same misconception about the alleged compulsive nature of bizarre desires is criticized in the Durham

---

[1]Ibid., p. 282. [2]Ibid., p. 282.

Rule's rejection of the "Irresistible Impulse" insanity test. That test "carries the misleading implication that 'diseased mental conditions' produce only sudden, momentary or spontaneous inclinations to commit unlawful acts."[1] Mental illness can produce deep, brooding depressions which gradually lead the insane person to a well thought-out crime.

The fact that bizarre, crazy desires are not necessarily irresistible highlights another problem for any formulation of the insanity defense: the need to distinguish between (in the circumstances) irresistible desires, and desires which are not in fact resisted. The most dangerous consequence of failing to make this distinction is the attempt to use the mere fact that someone commits a bizarre crime as evidence that he must have been subject to uncontrollable urges. (We may call this "the problem of self-excusing crimes.") Clearly it is undesirable to rule out a priori the possibility that someone commits such a crime, yet could and should have refrained from doing so. The opinion in U.S. v Brawner rules out such abuse of the insanity defense in the case of self-excusing crimes, in its "caveat paragraph":

---

[1] Durham v United States, (U.S. Court of Appeals, D.C. Cir., 1954), Section II.

> The terms "mental disease or defect" do not include an abnormality manifested only by repeated criminal or otherwise anti-social conduct.[1]

This problem is particularly acute in the case of "mental illnesses called character or personality disorders," the main symptom of which is repeated anti-social behavior.[2] What is needed is independent evidence of mental illness, and of how it impaired the rational behavior control of the allegedly insane criminal.

What follows for my criterion of responsibility? Does it provide no way to distinguish between the culpable, rational wrongdoer, and the blameless, insane victim of abnormal desires? And within the class of insane people, does it give us no means of deciding who could, and who could not, be expected to resist her desires? The crucial question, according to my criterion, is whether they were capable of basing their actions on morally justifiable qualitative evaluations. The preceding arguments suggest that this question cannot be answered, since the mere fact that a desire is abnormal does not make it irresistible. However, all is not lost. H.L.A. Hart points out that in our everyday ascriptions of moral responsibility we are perfectly well able to distinguish the person who should have tried harder to resist temptation from the person who,

---

[1] U.S. v Brawner, "Comments," Section 6e.

[2] See Gaylin, op. cit., pp. 250-1.

try as she might, could never resist. With knowledge and experience, we can extend this confidence to our judgments as to when mentally ill people may be expected to resist their anti-social urges.[1] These everyday intuitions can provide the basis for judgments as to when someone could have exercised qualitative evaluations. Moreover, even if a clear-cut dividing line cannot be drawn between the rational wrongdoer and the blameless mentally ill person in terms of the resistibility of their desires, Feinberg has already provided a justification for treating the latter more leniently. As explained above, unlike the rational criminal, the mentally ill lawbreaker often gains nothing from his crime.

However, even Hart admits that in many insanity defense trials the resistibility issue is very hard to decide. Evidence based on psychiatric evaluations of the accused is only indirectly related to his state of mind at the time of the crime.[2] I think the same difficulty can arise in hard cases involving perfectly sane people. Was I able to resist the sexual advances of the beautiful married woman, to whom I had long been attracted, after we had both been drinking freely at a party? Was I able to act on my

---

[1] H.L.A. Hart, "Changing Conceptions of Responsibility," op. cit., p. 203.

[2] Ibid., pp. 204-5.

qualitative evaluation that, all things considered, I ought to reject her suggestions? A prisoner of war in a Nazi concentration camp is told that, if she helps in the selection of prisoners for execution, she and her children will be spared. She knows that collaborating with the Nazis is a terrible sin, but can she be blamed for not acting on this qualitative evaluation? It must be admitted that my criterion of responsibility gives no determinate answer in these cases. The point of these examples is to show that our intuitions are uncertain not only in tricky insanity cases, but also in some of our assessments of sane, rational people. The question of whether one could have acted on one's qualitative evaluations is partly normative, and inevitably involves discretion.

My defense of my criterion is this: the fact that it is difficult to apply in hard cases in no way discredits it. This merely reflects the nature of the hard cases, and indeed we should be suspicious of any criterion of responsibility that produces an easy answer to them. The purpose of a standard of responsibility is to reflect and justify our considered intuitions in particular cases. I claim that my criterion does indeed follow our intuitions in clear cases; and that it reflects the complexity of hard cases, while providing sound guidance as to how we should go about deciding them. Deciding individual hard cases of responsibility will require, among other things, detailed

empirical evidence about the agent's mental state and circumstances. My concern has been to show that there are no conceptual barriers to making reasonable attributions of responsibility, even in hard cases.

5. Hard Determinism Revisited

The fact that my criterion of responsibility appears to correspond rather well with our intuitions would not impress the hard determinist. The whole point of his argument is that, if we carefully follow out the consequences of determinism, we will realize that these intuitions are just mistaken. In fact, we may never justly be held responsible. At the outset of this chapter, I stated that I would assume for the sake of argument that compatibilists have successfully proven the possibility of moral responsibility on at least some occasions. Now it is time to return to the hard determinist objections that were temporarily suspended, and in particular to see how they apply to the criterion of responsibility sketched in this chapter.

The account of self-determination given in Section 3 of this chapter is designed to forestall the hard determinist image of us as the passive observers of the causal factors which determine our actions. These causal factors are our beliefs, desires, and character, which are themselves shaped by our family, social and cultural environment. The

importance of our forming and acting on qualitative evaluations is that, in contrast to the hard determinist picture, it enables us to take control of and shape our desires and character. These evaluations, moreover, are based on judgments of worth according to criteria which we ourselves create.

However, this account of self-determination seems to be vulnerable to the same hard determinist objection raised by Edwards, quoted by Dennett in Elbow Room (pp. 83-4). Consider this example, similar to the one discussed by Edwards. Two women, with very similar personalities and from exactly the same social background, have an unfortunate tendency to become possessive and jealous in their relationships. Both of them form the qualitative evaluation that this is a self-destructive and selfish tendency. To rid themselves of it, they both discuss it with friends, read self-help books, and tell themselves how undesirable jealousy is. However, only one of them succeeds. How would we judge the woman who fails, according to my criterion? We may decide, on studying her character in depth, that certain deep-seated neuroses made it impossible for her to succeed. But it is surely also conceivable that we would decide that she could have succeeded had she tried harder, and is hence responsible for her failure. We would point to the fact that the other woman, who is very similar, did overcome her jealousy. In

general, we want to leave open the possibility of people who try to improve themselves and fail, but could have succeeded had they tried harder.

This is where the hard determinist attacks. The unfortunate woman, he would argue, did everything right in trying to improve herself. Her failure to put forth the extra effort which would have helped her overcome her jealousy was itself determined by genetic and environmental factors beyond her control - factors with regard to which the successful woman just happened to be dealt a better "hand." Is it fair to blame the former for failures over which she ultimately had no control? Similarly, the wrongdoer who does not even <u>try</u> to improve himself lacks the moral strength to do so simply because of a bad genetic and environmental "hand." In this way, any failure to form or act on qualitative evaluations can be passed off as the blameless result of causal factors beyond the agent's control. My account of self-determination thus does not avoid the hard determinist argument, but instead postpones it from our actions to our meta-level self-control (i.e., our acts, or omissions, of qualitative evaluation).

Underlying the hard determinist argument is the principle of alternate possibilities (PAP): in order to be responsible, one must have been able to do otherwise. Given their causal antecedents, the hard determinist argues, people who fail to act on justifiable qualitative

evaluations could not avoid failing. At this point, we need to examine the arguments given by compatibilists in rejecting the PAP. Such arguments, if successful, will neutralize the hard determinist argument. Both Frankfurt and Dennett would admit that, in the exact circumstances in which she found herself, the woman who failed to overcome her jealousy would always fail. However, they would claim that this is irrelevant to her moral responsibility.

In his innovative attack on the PAP,[1] Frankfurt considers cases where a person's actions are over-determined. Thus a man decides of his own free will to commit, say, a murder, which he carries out. However, unknown to him, a malevolent third party wants him to commit the murder. If the murderer had for any reason changed his mind, the third party would have intervened to cause him to commit it (by hypnosis, brain surgery, drugs, or some other infallible method). Because the third party is waiting in the wings to intervene should the murderer change his mind, the murderer could not do otherwise. Yet, Frankfurt plausibly suggests, we would still hold him responsible for the murder, proving that the PAP does not hold in at least one type of case. What matters is that it was his own deliberation that led him to kill. The impending intervention of the third party played no causal

---

[1] "Alternate Possibilities and Moral Responsibility," Journal of Philosophy, Dec. 1969.

role in the murder, which would have happened even if the third party did not exist.

Unfortunately, Frankfurt's examples are all rather fanciful, and, as Dennett points out, supporters of the PAP might "try for a patch" (Elbow Room, p. 132), and insist that the PAP holds except in rather bizarre cases of over-determination. However, Frankfurt also argues against the PAP in more everyday cases. This occurs in a slightly cryptic, compressed section at the end of his article "Freedom of the Will . . .," the rest of which is discussed earlier in this chapter.[1] Frankfurt's discussion is in terms of second-order volitions to act on first-order desires (i.e., to make them our will), but it can easily be translated into my notion of qualitative evaluation. Suppose that someone, under no compulsion, willingly acts on one of his qualitative evaluations:

> Even supposing that he could have done otherwise, he would not have done otherwise; and even supposing that he could have had a different will, he would not have wanted his will to differ from what it was. Moreover, since the will that moved him when he acted was his will because he wanted it to be, he cannot claim that his will was forced upon him or that he was a passive bystander to its constitution. Under these conditions, it is quite irrelevant to the evaluation of his moral responsibility to inquire whether the alternatives that he opted against were actually available to him.[2]

---

[1] See Frankfurt in Watson, op. cit., pp. 93-5.

[2] Ibid., p. 94.

Above, Frankfurt argued that the fact that the murderer freely chose to kill renders irrelevant the fact that, in any event, the malevolent third party would have ensured that he committed the murder. Similarly, in the passage quoted he is saying that even if (as hard determinists insist) the man could not have made and acted on any other qualitative evaluation than the one he actually made, his responsibility is not negated. Crucially, even had he been able to act on a different evaluation, he would not have wanted to do so. By voluntarily making qualitative evaluations (second-order volitions, in Frankfurt's terms) one thus takes responsibility for one's actions, and renders irrelevant the question of whether one could have avoided making that evaluation.[1]

If successful, Frankfurt's argument will rebut the extended hard determinist challenge. It may be true that ultimately our ability to act on qualitative evaluations is caused by factors beyond our control. Hence, given our precise mental state, current circumstances, and past history, only one qualitative evaluation is open to us. Nonetheless, Frankfurt argues, provided that we freely made that evaluation, we are responsible for it. Frankfurt's analysis is particularly plausible when applied to his example of the "willing addict." This person is physically

---

[1]This is a further application of Frankfurt's notion of "identifying oneself decisively" with a desire, and thus making it one's own. See Ibid., p. 91, and my discussion in Section 3 of this chapter.

addicted to a drug and is unable to stop taking it, even if he wanted to stop. He is unable, that is, to act on qualitative evaluations that taking the drug is bad. However, he is "altogether delighted with his condition,"and even if he were able to escape from his condition, he would not want to. In fact, he would take steps to re-acquire the addiction. In these circumstances, the irresistible nature of his addiction does not in the least diminish his responsibility, since he welcomes it and would still choose to be addicted even if he were able to escape it. His desire for the drug, and not the fact that he cannot avoid taking it, best explains why he takes it. (Similarly, the murder in an earlier example is best explained by the murderer's desire, and not by the impending intervention of the third party.)

However, Frankfurt's account is ill-suited to my example of the woman who wants to rid herself of her possessiveness and jealousy. The fact that (as the hard determinist would insist) she was unable to rid herself of possessiveness <u>is</u> relevant to her responsibility. If she had been able to rid herself of it, she would have done so: in fact she <u>tried</u> to, but failed. Her possessive behavior is best explained not by her desires, but rather by other causal factors which operate <u>against</u> her desires, or at least against her qualitative evaluations. In general, cases in which people struggle unsuccessfully to

change themselves are less amenable to Frankfurt's argument against the PAP and the hard determinist challenge. In these cases, the inability to do or choose otherwise does seem to threaten moral responsibility.

Dennett's attack on the PAP is based on examples which are more everyday than either of Frankfurt's groups of cases. As mentioned above (Chapter II, Section B4), Dennett points out that in many cases being unable to do otherwise is a highly commendable attribute. Nearly everyone exhibits this attribute every day, by not committing gratuitous murders, rapes, muggings or other atrocities. Furthermore, we would claim that nothing could make us do these things. Despite this fact, this inability is considered an admirable, emulable quality, even though for most of us it is an effortless consequence of being brought up by non-psychopathic parents and teachers. On other occasions, we have to work at achieving self-control. Thus a child abuser may sincerely regret his actions, enter a rehabilitation program, and afterwards be genuinely incapable of repeating such crimes. Whether our self-control is due to our moral education, or to our own efforts, it is morally commendable. (However, it would of course be strange to actually praise someone for such routine self-control, since we expect this minimal level of decency from all competent people.) Being unable to do otherwise hence does not always rule out responsibility.

Dennett's argument has the salutory effect of shifting our attention to situations where there is no serious moral conflict. Incompatibilists have made the mistake of focusing on tricky moral dilemmas, in which our decision is in the balance.[1] They ignore the fact that we are often praised, or at least highly regarded, precisely because, thanks to our good moral character, no dilemma arises for us.

I believe that Frankfurt's and Dennett's attacks conclusively refute the PAP. They both produce convincing counter-examples, involving people who experience no moral conflict, and happily and effectively act on their qualitative evaluations. So many everyday counter-examples can be generated from their arguments that it would be heroic, but highly implausible, to try to salvage the PAP as a condition of responsibility by further modifying it. However, in other cases the PAP does retain its plausibility, which indicates that though the principle may not be universally true, the hard determinist threat is still alive. These cases involve people who, like the possessive woman in my example, struggle unsuccessfully to improve themselves. As I argued above, Frankfurt's argument fails to prove that, in these cases, whether one

---

[1] e.g., C.A. Campbell (op. cit.), who argues that moral responsibility only arises when there is a conflict between our strongest desire, and what we perceive to be our moral obligation. (See above, Chapter I, Section 4a(i).)

could have done otherwise is irrelevant to moral responsibility. And while Dennett shows great insight in pointing out everyday cases in which the PAP does not hold, he should not reverse the error of his opponents by ignoring cases of moral struggle, where the PAP does seem to apply. These considerations also present problems for my own account, which locates responsibility in our ability to define ourselves by means of qualitative evaluations. These areas of moral struggle are precisely those where the hard determinist argument is hardest to dispel.

In general, the hard determinist argument seems to retain its force in the case of malevolent or otherwise untoward actions, whether or not moral struggle has occurred. The hardened bank robber commits robberies willingly, knowing that they are wrong. Is it irrelevant to his responsibility for the hard determinist to insist that, given his background and psychological make-up, he could do no other? In all instances of malevolent actions, and failures to improve oneself by acting on qualitative evaluations, the hard determinist closes the gap between the actual and the possible. Given her precise history and character, the possessive woman could not have succeeded in improving herself, any more than the bank robber could have avoided his life of crime. We need to examine Dennett's other replies to the hard determinist challenge.

The argument from Dennett which we have considered questions the truth of the PAP. His other arguments are more pragmatic. Even supposing the PAP to be true, Dennett argues that in practice it would be impossible to know whether anyone was responsible. (See above, Chapter II, Section B4.) Furthermore, the question of whether we could have done otherwise in exactly the same circumstances - which, as compatibilists such as Mill fully admit, is ruled out by determinism - is of no practical value. The forward-looking, utilitarian view of blaming and responsibility advocated by Dennett is concerned only with projectible misdeeds. These are misdeeds which we are likely to repeat in similar circumstances, which may well arise in the future. The goal of preventing further wrongdoing will not be advanced in the least by knowing what we would do in exactly the same circumstances, since these micro-circumstances will never again arise. Moreover, overlooking the "micro-details" (Elbow Room, p. 162) of the circumstances surrounding a misdeed will have a most beneficial consequence. Rather than using the always-available causal explanation to excuse their misdeeds, by showing that they could not, in the circumstances, have done otherwise, people will be encouraged to take responsibility for their actions. They will realize that other people will not accept such excuses, and hopefully they will begin to hold themselves

accountable for their misdeeds. Such a consequence is clearly healthy for all of society.

The rest of Dennett's argument consists in challenging what has been conceded until now: namely, that the truth of determinism does indeed prevent us from doing otherwise. If Dennett's challenge is successful, then even if the PAP survives his earlier attack, determinism does not rule out responsibility. Dennett does not offer a hypothetical analysis of "could." Instead, he asks what sense of "being able to do otherwise" is needed to provide "elbow room" between the actual and the possible, and hence to allow for moral responsibility. The required sense is provided by the notion of epistemic possibility. Even if all of our actions are determined, we are able to do otherwise, in the sense that alternative actions are epistemically possible. Even though only one future course of events is physically or "nomically" possible, so far as we or any other finite beings can ever tell, several alternative paths remain open. The epistemic openness of the world - i.e., the fact that certain processes, including some that cause our actions, will always be unpredictable by us - is what gives point to our deliberations over how we should act. Even though the die is cast, so to speak, by the laws of nature, to the best of our knowledge more than one course of events and actions always remains possible. For more detail on these points,

see my discussion of Dennett above, in Chapter II, Sections B3d and B4.

There is a strong suspicion that, despite the ingenuity of his argument, Dennett is trying to glorify our ignorance. It is paradoxical that our human limitations can be cited as necessary conditions for our freedom and responsibility, when elsewhere in his book Dennett explains how knowledge enhances our control of our lives (Elbow Room, Chapter 3). In fact, there is a venerable tradition in philosophy which regards knowledge as enhancing our freedom.[1] We can grant that Dennett's pragmatic reasons for believing that and acting as if we have genuine options are conclusive. Even if we were to become convinced of the truth of hard determinism, it would still be rational to strive and deliberate as best we can, thus making the best of a severely restricted situation. However, whether people really can act otherwise than they do, and whether it is justifiable for us to hold them responsible, are separate questions from what it is prudent for us to do. The suspicion that Dennett is asking us to take the irrational course of basing responsibility upon ignorance is fuelled by an admission he makes. According to our "scientific image," he concedes, only one course of action

---

[1]Consider Plato's ideal of the man whose desires and passions are ruled by his reason; and Kant's view of the free man as being subject only to the dictates of reason. (However, Dennett criticizes the latter for making unreasonable "absolutist" demands. See Elbow Room, p. 49.)

is nomically possible for us on any given occasion. The "more fine-grained standpoint" (Elbow Room, p. 143) provided by the scientific image shows that, contrary to our perceptions, only one future is "really" open. Doesn't this scientific image reveal that our "manifest image," upon which we base our belief in alternate possibilities, is just a myth, albeit a useful one?

To his credit, Dennett does deal directly with the objection that epistemic possibility does not constitute the "real, objective possibility" of doing otherwise (Elbow Room, p. 151). His reply is that there is no gulf between our imperfect knowledge, and the hard, objective truth about how we are determined to act. Because of the existence of "unstable, chaotic systems," the world itself, and not merely our limited perception of it, is indeterminate. Only an infinite being would be able to predict with certainty everything that we do, and even such a being would be incapable of it in the case of quantum-indeterministic processes. Is it reasonable, asks Dennett, to bemoan our ignorance in comparison with an infinite being? The answer is clearly "no": it would be irrational to yearn for what is logically impossible for us finite beings. However, Dennett has merely changed the subject. The objection to his argument does not even concern our ignorance; rather it concerns the implication of the existence of sufficient causal explanations for our

actions, even if only an infinite being could know in advance of these explanations. In these circumstances, is it fair or just to hold people responsible for their actions, and for their attempts at self-determination? Dennett's pragmatic arguments do not even touch on this moral question.

Dennett would doubtless reject the dichotomy just presented between pragmatic and moral considerations. Especially in the final chapter of Elbow Room, he adopts a utilitarian approach to attributions of responsibility, tending to replace questions of justice by determinations of how to produce the best consequences. In the final chapter, I intend to argue for a retributive theory of punishment, and to show how it is consistent with Dennett's and my own compatibilist arguments. For the time being, I will merely point out that Dennett's wholly utilitarian approach does not even provide a vocabulary in which to express the moral doubts, considered here, concerning the implications of determinism for moral responsibility. Since these doubts are substantial, this fact alone counts against a purely utilitarian view.

The moral doubt which, I have argued, is not addressed by Dennett is this: given the existence of sufficient causal explanations of all human actions, including our attempts to control our own destinies by means of qualitative evaluations, is it fair to blame people for their misdeeds? The person who fails to improve herself,

yet whom we deem capable of succeeding had she tried harder, is responsible according to the criterion developed in this chapter. Yet if this failure is itself caused by genetic and environmental factors, isn't she just unlucky that she wasn't dealt a better "hand" with regard to the motivational factors which prevented her from succeeding? In other words, the hard determinist argument has led us to the problem of moral luck. Dennett's pragmatic arguments give us excellent reasons for continuing to deliberate to the best of our ability; to try to improve ourselves by questioning and shaping our goals, character, and values; to punish people for their misdeeds; and to hold ourselves and other people accountable for their actions. However, they do not dissolve the stubborn doubt that ultimately, in a sense not expressible in a purely utilitarian vocabulary, our practice of holding people responsible is unjust.

6. Summary

In this chapter I have presented a criterion of responsibility based on the control we may exert over our actions, character, and values. Such an analysis is needed to forestall objections to the traditional compatibilist view that it ignores the compulsion that can be exerted by our own desires. My account of how we can acquire responsibility by means of self-determination is intended to complement Dennett's arguments. I believe that in Elbow Room, he gives a powerful general defense of compatibilism;

but that it needs to be supplemented by a positive criterion to determine exactly when a person is responsible, which is what I have presented. To support my criterion, I have shown that it meshes rather well with our intuitions concerning people who are traditionally excused from responsibility.

However, my account of responsibility is vulnerable to a persistent application of the hard determinist argument. I then turned to Frankfurt and Dennett's arguments against hard determinism. Though they successfully refuted the PAP, I concluded that they did not deflect the main thrust of hard determinism. In particular, I examined the responses of Dennett, whose work this chapter is aimed at expanding and improving. I argued that in his own pragmatic, utilitarian terms, Dennett's replies are conclusive. However, he does not dispel the nonconsequentialist argument that, because of the existence of moral luck, it is unfair to hold people responsible.

In the next chapter, I will discuss whether Dennett successfully answers these nonconsequentialist objections in his rather brief discussion of moral luck. Since I myself will, in opposition to Dennett, advocate a nonconsequentialist, retributive view of punishment (in Chapter V), it is particularly important for me to solve the problem of moral luck. Unlike Dennett, I cannot shrug it off as caused by an unrealistic "absolutist" demand, and solve it pragmatically.

CHAPTER IV

MORAL LUCK

The problem of moral luck is the final battleground for the conflict between my compatibilist account and the hard determinist challenge which has proven to be its greatest threat. I have endorsed Dennett's naturalistic account of human agency, which is intended to show that rationality, control, self-determination, and effective deliberation are all possible, even if our actions have sufficient causal explanations. These necessary conditions of responsibility have been shown to withstand hard determinist arguments. However, even this impressive description of human agency, supplemented by my detailed analysis in Chapter III of the structure of self-determination, is vulnerable to an extension of the hard determinist argument. No matter how versatile, ingenious, creative, and strong-willed we are in our attempts to control and improve ourselves, we are limited by factors beyond our control. Due to differences in genetic endowment and environmental ease, some people are simply less capable of healthy self-determination than others. These differences are a matter of luck: moral luck. For my defense of compatibilism, it is not "can,"

but rather moral luck, which is the "frog at the bottom of the beer mug,"[1] defying dissolution by my arguments.

1. The Threat Posed by Moral Luck

There are two well-known accounts of the destructive consequences of moral luck for our concept of moral responsibility, both called simply "Moral Luck."[2] It will be useful to consider these pessimistic arguments before turning to solutions.

Nagel regards the problem of moral luck as the natural consequence of taking seriously the requirement that, in order to be responsible, we must be able to control our actions. As we saw in Chapter III, sections 4a and 4b, ignorance and compulsion excuse because they prevent a person from controlling his actions by his deliberations. Actions, or their outcomes, are, by definition, beyond our control to the extent that they are subject to luck. Nagel argues that the effect of moral luck in determining our actions is ubiquitous, and hence that our control over all of our actions is undermined.[3]

---

[1] J.L. Austin, "Ifs and Cans," Philosophical Papers (Oxford, 1961), p. 179, quoted by Dennett, Elbow Room, p. 19.

[2] Thomas Nagel, in Mortal Questions (Cambridge, 1979); and Bernard Williams, in Moral Luck (Cambridge, 1981).

[3] Nagel, op. cit., p. 26.

He considers four ways in which our actions are "disturbingly subject to luck." 1) "Constitutive" luck, i.e., the kind of intellectual, emotional, motivational, and physical endowments with which we happen to be born. 2) "Luck in circumstances," i.e., the kind of environment we grow up in, and the kind of situations it is our (mis)fortune to meet throughout our lives. For instance, lucky indeed is the person who, unlike the woman in my example in Chapter III, only ever gets involved with devoted, adoring lovers, and is never given the slightest cause for jealousy. 3) "Luck in how one is determined by antecedent circumstances."[1] 4) Luck in the way our actions turn out. A notorious example of this is the drunken driver who hits and kills a child at a pedestrian crossing. Another driver, equally drunk, careless, and incapable of stopping should a child use the same pedestrian crossing, is fortunate enough to reach home without encountering any children.

First I have a minor criticism of Nagel's classification of types of moral luck. It is not clear why he needs the third category of moral luck, nor exactly what it consists in. Surely how one is determined by antecedent circumstances is a function of the first two kinds of moral luck: one's genetic endowments, and the situations one encounters. According to the definition of determinism

---

[1] Ibid., p. 28.

proposed by Van Inwagen (see Chapter I, section 4c) how one turns out is a consequence of the state of the world at earlier times (which includes one's genetic structure and environment), and of the laws of nature. I propose, then, to consider only three sources of moral luck which are clearly distinguishable chronologically: 1) our genetic endowments; 2) our environment, and the situations we encounter throughout our lives; and 3) the effects of luck after we have acted. Following Nagel, I call the first two kinds constitutive and circumstantial luck; and I dub the third kind as "consequential luck."

Nagel starts off with familiar cases, most of them of bad moral luck: the person born with a bad character, who has to struggle unusually hard to avoid misdeeds (constitutive); the Nazi war criminal who, had he chanced to be born in another time or place, would have lived a "quiet and harmless life" (circumstantial); and the two drunken drivers (consequential). These examples elicit our sympathy, but it may be maintained that these are extreme cases which leave intact our responsibility in everyday cases. However, Nagel's point is that these extreme cases give rise to a slippery slope: all of our actions are subject to moral luck, even though less dramatically than in the extreme cases. The fact that moral luck infects the past, present and future influences on our actions and their outcome undercuts - or at least diminishes - our

responsibility in all cases. Moral luck is not an occasional anomaly, but rather an essential feature of the human condition.

What turns moral luck into a paradox is the fact that our reflective moral judgments do not compensate for its effects. This simply violates the belief that control is necessary for moral responsibility. Like it or not, we do view murder as more serious than attempted murder, even if the victim's (not) wearing a bullet-proof vest is the only difference between the cases.[1] The decision to instigate a violent revolution, with all the costs it will involve, will be justified only if it succeeds in setting up a more peaceful, humane society; and this depends in part on factors beyond the leader's control.[2] People are praised or blamed for character strengths or flaws, which are squarely a matter of constitutive luck.[3] Our everyday judgments involve in part the "moral determination by the actual." Regardless of the lack of control we can exert over the influence of moral luck, we do not judge people solely on the basis of what they can control.

> We judge people for what they actually do or fail to do, not just for what they would have done if circumstances had been different.[4]

This cannot be squared with the deference we do pay to the requirement of control when we recognize the excuses of

---

[1] Ibid., p. 29.

[2] Ibid., p. 30.

[3] Ibid., pp. 32-3.

[4] Ibid., p. 34.

ignorance, compulsion and insanity. These excuses seem to be firmly based on a Kantian view of responsibility: we are responsible only for our will, or intentions, which are immune to the vagaries of good and bad fortune. To make our moral judgments consistent with the Kantian view, we need to eliminate the role of luck. Thus the person who attempts murder will be just as morally "stained" as the murderer, and people who unluckily cause harm will be excused. The object of our judgments will be people's intentions, and not the outcome of their actions.

The problem with this move is that, assuming the truth of determinism, these intentions are just as subject to (un)lucky causation as are actions themselves. Joel Feinberg, to whom Nagel briefly refers, has shown how one's intentions may be unluckily triggered off by chance events, or extreme provocation, both of which are beyond one's control.[1] This is, indeed, an inevitable consequence of the truth of determinism and the existence of constitutive and circumstantial luck. If we persist in trying to isolate an "inner core" of our actions which is immune to moral luck, and is hence the proper object of our moral judgments, we will arrive at "an extensionless point." Indeed, says Nagel, the consequence of determinism is "the absorption of [the self]'s acts and impulses into the class

---

[1] Joel Feinberg, "Problematic Responsibility in Law and Morals," Doing and Deserving (Princeton, 1970), pp. 34-7.

of events."[1] The lack of an object for our judgments of responsibility is a particular consequence of the thesis that determinism seems to rule out the existence of a self at all.

Here is Nagel's account of the dilemma of moral luck: our Kantian intuitions require that we restrict our moral judgments to factors not subject to luck; but our actual, considered judgments violate this requirement. Moreover, revising our judgments to make them consistent with our Kantian intuitions would leave no one or nothing for us to judge.

Williams focuses on consequential luck. In particular, he considers Gauguin and Anna Karenina, both of whom left their families to pursue other goals: Gauguin, his life as an artist, and Karenina, her romance with Vronsky. Rather than directly examining our moral intuitions about their actions, Williams considers Gauguin's and Karenina's likely attitude toward their own actions. In what circumstances are they likely to view their decisions to leave their families as rationally justified? Williams does not spell out the connection between their subjective evaluations of their actions, and our moral evaluations of them. However, it is reasonable to assume that, for example, Gauguin's finding his decision to be rationally justified is necessary for it to be morally justifiable.

---

[1] Nagel, op. cit., p. 36.

William's main contention is that Gauguin's action will only be justified to himself if he succeeds in developing his talents as a great artist. The risk, or rather the certainty, of the harm caused to his family can only be justified by his success as an artist. The problem of moral luck arises because his success depends on two kinds of luck. First, he needs extrinsic luck, in that he must avoid accidents which would physically prevent him from painting. Such luck is not relevant to his self-evaluation, since even if it caused him to fail, the rationality of his original decision would not be impugned. Second, he must enjoy good luck in the development of his talents. This second kind of good fortune is intrinsic to his project. If he fails simply because, despite the lack of incapacitating accidents, his talent does not develop, his original decision will be "unjustified," and the failure will be his. In intrinsic failures, the entire project is shown to have been misconceived.

But surely Gauguin's success lies in his own hands, and can hardly be attributed to luck. Williams concedes this. Gauguin's luck consists rather, he argues, in the fact that his "gamble" on his future success paid off. However confident and well-supported was his expectation of success, the artistic process is so complex and delicate that it is very unlikely that, at the time of his decision,

he knew that he would succeed. He has thus benefited from epistemic luck. Even if we decide, in retrospect, that he did indeed know that he would succeed (and this is unlikely in Gauguin's case), luck remains in the fact that the right considerations occurred to him prior to his decision.[1] The role of luck in most other decisions which require "retrospective justification" is plainer to see. For instance, whether Anna Karenina will feel justified in leaving her family will depend in part on whether Vronsky continues to love her. His feelings are clearly, to some degree, not up to her. Similarly, to take Nagel's example, the bloodshed caused by a revolution can only be justified by its success in creating a better society. Whether it succeeds depends on a complex of political factors, many of which are beyond the leader's control, and unpredictable by him. What is crucial to the previous two examples is that factors beyond the agent's control are intrinsic to his project's success. It follows that whether he succeeds is partly subject to luck. The value of the Gauguin example is that it provides a limiting case, in which whether his decision and action are justified or unjustified is largely within his control. If even this case is, as Williams argues, subject to intrinsic luck (epistemic luck, in this case), then it will be easy to construct many other cases

---

[1] For Williams' account of extrinsic vs. intrinsic luck, and the concept of epistemic luck, see Williams, op. cit., pp. 25-7.

in which the justification of an action will be more obviously subject to intrinsic luck. The Anna Karenina and revolution examples are two such cases.

All of these examples are cases of "determination by the actual" (cf. Nagel).

> One's history as an agent [which includes one's attitude to one's own actions] is a web in which anything that is the product of the will is surrounded and held up and partly formed by things that are not.[1]

This is the result of the dependence of the justification of these decisions and actions on future contingencies, i.e., retrospective justification. This threatens the objectivity and importance of rationality itself. Surely the reason we prize rationality is precisely that it is supposed to be immune to luck. A decision may, by an unlucky coincidence, produce a bad consequence; but, provided that it was made carefully in the light of all the available evidence, we need not declare it irrational. Rationality is supposed to be immune to luck in the same way as a Kantian good will.

Maybe too much has been conceded to Williams too fast. Isn't it possible to make an <u>independent</u> judgment of the rationality of a person's decision as "reasonable or unreasonable relative to his situation, whatever its actual outcome"?[2] Thus we could, for example, apply a game-theoretic analysis to Gauguin's action in leaving his

---

[1] Ibid., p. 29.

[2] Ibid., p. 31.

family. By assigning probabilities and values to all possible outcomes, we can judge his action as rationally justified if it maximizes the "expected utility." Such judgments would not be subject to moral luck.[1]

Williams admits that we frequently do make such judgments about the "intrinsic rationality" of our decisions, as part of our forward-looking concern with our rationality on future occasions. A perfectly rational decision may unluckily lead to a bad result, yet give us no reason to alter our deliberative strategy in the future. As Dennett points out in his discussion of "designing a perfect deliberator," even an optimal decision procedure will not be infallible.[2] Similarly, we may have good reason to regret our decision, or at least to change our future strategy, if it is poorly conceived and only succeeds because of good luck. Because of the greater destructive consequences for the future of a faulty reasoning process, as opposed to a single mistake or stroke of bad luck, "a more fundamental form of regret is directed to deliberative error than to mere mistake."[3]

Unfortunately, Williams argues, in the kind of case under consideration in his article, such "pure" assessment

---

[1] Williams discusses the possibility of such assessments of the intrinsic rationality of decisions in some depth, Ibid., pp. 31-6.

[2] Elbow Room, pp. 139-44.

[3] Williams, op. cit., p. 34.

of deliberative rationality, isolated from its consequences, is not possible. The goal in the light of which a decision is deemed rational will only carry weight if it is in fact achieved. Before the goal is achieved, it can only be assessed according to the deliberator's current perspective, from which it has not yet acquired value. For example, it is only from the perspective of a successful artist that Gauguin's future can attain the value which justifies his decision; and, however likely or even certain his success was, this perspective was not available to him at the time of his decisions.

> . . . the project in the interests of which the decision is made is one with which the agent is identified in such a way that if it succeeds, his stand-point of assessment will be from a life which then derives an important part of its significance from that very fact; if he fails, it can, necessarily, have no such significance in his life.[1]

The upshot of this is that, in at least some cases, a treasured feature of rationality is absent. It is simply not possible to assess these decisions (and the actions which are based on them), in the light of all the evidence available at the time when they are made, as rational or justified. The final verdict on their rationality must await the outcome of the actions based on them, and is hence to some degree dependent on contingent matters beyond the agent's control, i.e., subject to luck.

---

[1] Ibid., p. 35.

How, finally, does this conclusion apply to moral luck, or indeed to morality at all? It must be said that Williams' account of this crucial, final stage of his argument is rather opaque.[1] The gist of his argument seems to follow from his claim that rationality has been shown to be to some degree subject to luck. Insofar as morality is regarded as being based on reason, it too will be subject to luck. (This assumes, what is surely uncontroversial, that being rational is a necessary condition for being morally responsible.) However, it has commonly been supposed that what gives moral (as opposed to practical or prudential) considerations their supreme value is precisely their status as rational decisions immune to the vagaries of chance. As Kant says, even if bad luck leads to a bad outcome, a good will "would sparkle like a jewel in its own right."[2]

This conclusion is not as destructive as Nagel's. Nagel thought that the existence of moral luck renders all of our moral judgments incoherent and unjustified; but that we are nonetheless committed to making them. Williams, on the other hand, seems to believe that morality can survive the existence of moral luck. However, it survives only in an attenuated form, losing the very feature which made it so important. Another respect in which Williams' arguments

---

[1] Ibid., pp. 36-9.

[2] Quoted by Nagel, op. cit., p. 24.

are less destructive than Nagel's is that, unlike Nagel, he does not claim that all moral appraisals are subject to luck. His conclusions only apply to cases of consequential luck; and among even these cases, only to those about which our judgments contain an ineliminable retrospective element.

2. Dennett's Reply

I have already discussed in some detail Dennett's solution to the problem of moral luck. (Chapter II, section B3c.) In this section, I will consider whether his solution is effective against Nagel's and Williams' pessimistic challenges. I concluded at the end of Chapter III that the pragmatic approach which Dennett adopts in the later chapters of Elbow Room leaves unanswered certain moral doubts about the justifiability of our attributions of responsibility. These doubts boil down to the problem of moral luck, so that, to the extent that his book is intended to give a non-pragmatic solution to the free will problem, it stands or falls with the section on moral luck. However, given the increasingly utilitarian drift as his book progresses, Dennett need not rest his case on the brief, principled (i.e., non-pragmatic) solution he gives to moral luck. Even if his argument fails, he could dismiss the whole problem as the product of unreasonable "absolutist" demands, and remain satisfied with a pragmatic

solution. I have argued that a purely utilitarian approach does not do justice to the depth of the free will problem; and in the final chapter I intend to show that compatibilism is consistent with a non-utilitarian view of punishment. It follows that I have to provide a non-pragmatic solution to the moral luck problem, hopefully in part by means of Dennett's arguments.

Dennett's arguments are particularly effective in dealing with Nagel, and indeed it is likely that many of these arguments were directed specifically at Nagel (though he is not named as their target). Nagel's dramatic claim that causal explanations of human actions would leave no room for agency at all seems to be squarely based on the DRA principle (determinism rules out agency).[1] It is against precisely this principle that the essays in Brainstorms (considered above, Chapter II, section A) were directed. Dennett's central claim in these essays is, as we saw, that physical descriptions of human actions are perfectly compatible with intentional ones. The richly detailed naturalistic account of agency which makes up most of Elbow Room (see above, Chapter II, section B3) further illustrates the compatibility of agency and causation.

The powerful and original move which is central to Dennett's treatment of moral luck in Elbow Room is to

---

[1] See my discussion of Norman Malcolm and Jennifer Trusted, in Chapter I, section 4b.

solution. I have argued that a purely utilitarian [illegible] does not do justice to the [illegible] of the [illegible] and in the [illegible] to [illegible] compatibilism [illegible] punishment. [illegible]

emphasize the role of skill in compensating for bad luck and in controlling our actions. Nagel makes the fatal mistake of denying any role to our skill, and regards all of our actions as the result of good or bad luck. Dennett points out, as we saw above, that the concept of luck only makes sense if it can be contrasted with actions that are not subject to luck.

It is true that Nagel is able to make a contrast between good and bad luck, or between good luck and its absence: thus he can contrast the beautiful with the plain, or the good-natured with the bad-tempered. However, this is not the contrast which Dennett requires to give meaning to the concept of luck; namely, the contrast with actions that are largely due to skill. We call games like Go-Fish "games of luck" to contrast them with "games of skill" such as chess, the result of which is largely determined by the players' ability. If skill never had an impact on any games, then the phrase "games of luck" would become redundant, and we would refer simply to "games." Similarly, when applied to human actions, "luck" provides a contrast with those factors that are within our control. If, as Nagel claims, "luck" is ubiquitous, and our skill never influences our actions, luck is once again a vacuous concept. Moreover, throughout Elbow Room Dennett gives ample evidence of the very real causal effect which our skill has on our actions. This evidence, and our everyday

experience, show Nagel's claim that luck is ubiquitous to be empirically false. When the role of luck is restricted in the way Dennett suggests, moral responsibility is no longer undercut. While Williams avoids Nagel's wilder claim that moral luck is ubiquitous, he too is vulnerable to Dennett's argument. In the few cases which he considers, Williams also downplays the degree to which, by the exercise of skill, we can control our actions.

In fact Margaret Urban Coyne has independently made a very similar point about moral luck. Attributions of moral responsibility only make sense if not everything is at the mercy of luck. Along with Dennett, she is not troubled by this, since she also holds that

> the possible contribution of luck to ventures is wholly compatible with the presence and exercise of rational choice, deliberation, wisdom, prudence, skill, foresight, calculation, practical reasoning, weighings of evidence, and so on.[1]

Indeed, Coyne goes on the offensive and suggests that, contra Nagel, far from ruling out agency, luck only makes sense in conjunction with agency. The meaning of luck is fixed by reference to agents' goals and expectations, in that good and bad luck unexpectedly further or thwart their goals.

It will be useful to examine in some detail the nature of the "skill" which Dennett believes can counteract moral

---

[1]Margaret Urban Coyne, Moral Luck?, paper presented at Eastern Division APA meeting, Boston, December 1983, p. 6.

luck. He illustrates the notion of skill by the use of sporting examples, but gives little detail on the corresponding kind of skill which is needed for moral responsibility. Whatever its exact nature, it is safe to assume that he believes that it is this skill which we exercise in the process of self-definition which he describes (and on which I expand in Chapter III, above).

Before we examine its details, it may seem that the notion of "moral skill" falls foul of a criticism forcefully made by Ronald Suter: there are no moral experts.[1] The existence of a skill seems to imply that someone with great proficiency in that skill would be regarded as an authority, in this case a moral authority. However, as Suter points out, this is a non-sequitur. Viewing moral goodness as a skill certainly implies that some people are better at it than others, and that we can all improve our "moral performance." Both of these implications are reflected in our everyday intuitions. The mistake is to assume that people who are generally regarded as exceptionally good people are thereby moral authorities. In morality, there are no universally accepted criteria by which we can determine who is an expert. This contrasts with other areas, such as carpentry, medicine, cars, etc., in which there are

---

[1]Ronald Suter, Are You Moral? (University Press of America, 1984), Chapter IV, pp. 49-60.

objective criteria to establish expertise, and where we willingly defer to expert opinion.[1]

Still, the notion of moral skill raises the sneaking suspicion that moral luck may re-appear in terms of varying degrees of endowment with moral skill. Let us look in more detail at how moral skill might be expected to counteract the effects, if any, of different kinds of moral luck.

a. Constitutive Luck

People are born with vastly different levels of intelligence, temperance, self-control, energy, perseverance, athletic ability, sensitivity, and many other qualities. Someone generously endowed with many of these qualities is likely to be morally skillful: being good will come more easily to such a person than to most people. Such differences in natural ability surely constitute a handicap to the moral progress of the less fortunate, and so even moral skill seems to depend on luck. Dennett's treatment of this kind of moral luck depends on his comparison of our moral progress with a marathon. Even substantial differences in our starting position will become insignificant over such a long distance. The process of self-definition which Dennett describes (and on which I elaborate in Chapter III) gives

---

[1]Ibid., pp. 56-7.

ample opportunity to overcome initial handicaps. Moreover, moral progress is not a competition at all, but the attempt to reach a base level of competence.

The marathon analogy is very successful in helping Dennett prove his main point that most people have ample opportunity to become morally competent, regardless of modest genetic endowments. Genetic handicaps do indeed restrict our intellectual and physical "raw materials," and put definite limitations on what career goals most people can realistically hope for; but being a morally responsible person is a minimal requirement which nearly everyone eventually meets. Infants, mentally retarded and insane people lack capacities which are essential for responsibility (see above, Chapter III, section 3). However, it is evident that most people do reach the minimal competence level required for membership in our moral community. Refraining from murders, robberies, assaults, rape, and other terrible crimes requires no great effort on our part. It need not be denied that those born with a generous and temperate character are more likely than others to perform supererogatory acts, and hence to earn extra moral credit. Furthermore, it need not be fatal to our moral scheme to admit that some people have to struggle to reach the threshold of moral competence. As Kant points out, such a person will deserve extra praise

for doing what comes effortlessly to most of us.[1] In this and other ways, our moral judgments already compensate for some of the effects of bad luck.

b. Circumstantial Luck

People's upbringing can vary just as dramatically as their genetic talents. The underprivileged child, who grows up in a poor, unloving, and violent environment has to overcome greater obstacles than most people do in order to achieve moral competence. Dennett's "marathon" analogy can be extended to cover such cases. Provided that these setbacks occur fairly early in life, one has plenty of time to use one's own efforts to compensate for them. (However, I will soon have to face the objection that a combination of severe genetic and circumstantial handicaps may raise insurmountable obstacles to moral competence.)

Luck with regard to the situations one has the good or bad fortune to encounter throughout one's life - life's lucky breaks and unfortunate setbacks - can be handled differently. It must be confessed that some people suffer from debilitatingly bad circumstantial luck: people whose houses are levelled by tornados, those who contract rare diseases, people laid off from work after 25 years' loyal service, etc. These cases do indeed present problems for

---

[1] Quoted by Nagel, op. cit., p. 33, footnote 8.

our moral evaluations of the victims of bad luck. According to the severity of the setback, these people can to some extent be forgiven for becoming bitter and resentful, and seem to deserve extra "moral leeway." However, this is conspicuously not the lot of most of mankind. For most of us, as Dennett points out, luck does indeed tend to average out in the long run. Furthermore, we can do much to minimize the impact that bad luck can have on our lives. An essential feature of being a rational deliberator, Dennett argues, is being able to isolate those features of our environment which are "chaotic." Realizing that these features are beyond our control and subject to luck, we can avoid committing ourselves to depending on any particular outcome.[1] In our long-term planning, we can avoid "putting all our eggs in one basket," and maintain back-up plans. On a more everyday basis, we can, for instance, take steps to stay healthy, and avoid dangerous situations, such as being alone in certain neighborhoods late at night. This is of a piece with Dennett's account of control in terms of foreknowledge. If we can anticipate factors which are out of our control, or at least be aware of the possibility of unexpected interventions, we can allow for such variables, and thus retain a good deal of control over our destiny.[2]

---

[1] Elbow Room, pp. 108-10.

[2] Ibid., p. 54-5.

our moral standards don't measure up to hers.

c. Consequential Luck

This kind of luck is similar to circumstantial luck in that it involves situations we encounter. It differs in that luck strikes subsequent to our actions, and that the outcome is a result, however unlucky, of our action. The same policy described above of intelligent anticipation of possible setbacks can be used to minimize the risk of bad consequential luck. This applies very straightforwardly to Nagel's example about the pair of drunk drivers. While the driver who kills the child can scarcely be called unlucky, the other driver is definitely lucky. Both drivers could have rendered their actions immune to at least this one piece of consequential luck by not drinking, and by driving more slowly.

Another consideration which restricts the importance of consequential luck is raised by Judith Andre. She discusses the "morally accident-prone" person[1] whose plans and actions always seem to end in disaster. Initially, each failure may be attributable to bad consequential luck. However, a long string of such failures should make us look for come character fault - rashness, clumsiness, an inability to foresee problems, etc. - which increases the probability of unlucky setbacks. (These character faults may themselves, of

---

[1]Judith Andre, "Nagel, Williams, and Moral Luck," Analysis 1983, p. 205.

course, be the product of bad constitutive or circumstantial luck.) In view of such morally accident-prone people, there are good prudential reasons for as much as possible avoiding attributing failures to moral luck. We want people to take control of their actions, rather than regarding themselves as the helpless victims of bad luck. Furthermore, we want to remove this catch-all excuse as a possible motive for self-deception with regard to the cause of people's failures.[1]

This argument is closely related to a pragmatic theme which runs through the later chapters of Elbow Room, and which is relevant to all three kinds of moral luck. It is always possible, Dennett admits, to seek "fine grained" causal explanations in the hope of excusing our behavior. Moral luck is an explanation which is particularly versatile and persuasive as an excuse. However, he argues,

> . . . healthy self-controllers . . . take responsibility for what might be, very likely is, just an "accident," just one of those things. That way, they make themselves less likely to be "accident" victims in the future.[2]

This gives us further prudential grounds for downplaying the role of moral luck in our lives. However, Dennett's solution to the moral luck problem does not rest on this pragmatic consideration. The latter is incidental to his

---

[1] Ibid., p. 207.

[2] Elbow Room, pp. 143-4.

courage, be the source of good consistently or
circumstantial (**ant.**) In vice or ever morally
sociopath-phone [soʊ-ʃ, sʌʊ-ʃ] **n.** a big good fundraiser to buy
for an [illegible] [illegible] of [illegible] [illegible]
word [wɜrd] **n.** [illegible] [illegible] [illegible]
[illegible]
[illegible]
[

main argument, which consists in emphasizing the role played by our skill in determining our actions.

It might be objected that Dennett's argument about the role of skill in our self-determination misses the crux of the problem posed by moral luck. Moral skill enables us, claims Dennett, to overcome the effects of bad constitutive, circumstantial, and consequential luck. But isn't this very skill itself dependent on our natural talents (constitutive luck), and perhaps also circumstantial luck? This is indeed the crux of the hard determinist objection to my account of self-determination (and hence of responsibility) considered in section 5 of the previous chapter. Any attempt to control our destiny will itself be subject to causation, which is beyond our control and subject to luck. Thus it was that the hard determinist argument, and the incompleteness of Dennett's pragmatic responses to it, ultimately boiled down to the problem of moral luck, which is the topic of this chapter.

Now, armed with Dennett's treatment of moral luck, I am at last in a position to refute the hard determinist argument. Rather than undercutting Dennett's solution, in terms of moral skill, to the moral luck problem, the hard determinist argument is a classic example of the very mistake which Dennett is criticizing. By attributing all of our actions, including our exercises of moral skill, to the effects of moral luck, it rules out a priori the

Diagram 7: One Shell in Active Air Ok Defense.

main argument, which concerned the cabbage and the rota

attributions of any actions to moral skill.[1] This is wrong for two reasons: (1) It renders meaningless the notion of (good or bad) luck, which makes sense only against the background of actions not attributable to luck (i.e., actions which result from our skill). (2) It violates our everyday experience, which is backed by Dennett's numerous, detailed examples, of the very real causal effect of our skill on our actions.

It is certainly logically possible that all of our actions are ultimately the product of luck. However, to describe them in this way is to stretch plausibility, especially when it is so much simpler to describe them as we usually do: the result of our deliberation, planning, creativity, skill, etc. Moreover, one might even grant the hard determinist his eccentric use of the term "luck," according to which everything we do is ultimately a matter of luck. Since what he calls luck is the very same thing that we ordinarily view as the result of our self-determination, we need not fear that luck, in his sense, rules out moral responsibility. The reason that luck, in its ordinary sense, is antithetical to moral responsibility, is that it signifies, by definition, the absence of these elements of our control over our actions. However, as I have argued, luck in this ordinary sense is

---

[1] See esp. Ibid., pp. 96-7, and my discussion of this, Chapter II, section B3c.

usually overcome by our moral skill. It is still possible that the hard determinist may come up with an argument to prove that all of our actions should be excused; but it is a mistake to try to do so by attributing all actions to good or bad luck.

Dennett's insistence on the efficacy of moral skill is entirely successful in refuting Nagel's sweeping claim that moral luck threatens all attributions of responsibility. Moreover, it is successful in refuting the final thrust of the hard determinist argument. Moral luck is not a "metaphysical" affliction undercutting everybody's responsibility. Even Williams' more modest conclusions about the destructiveness of moral luck for our moral judgments are vulnerable to Dennett's same argument. However, I need to consider substantial doubts which remain about particular instances of moral luck.

3. The Slippery Slope Revisited, Again

Unfortunately, Dennett runs the risk of losing the substantial gains he has made when he considers those people - "patzers at life"[1] - whom we routinely do excuse from responsibility. They include babies, young children,and all animals. They also include mentally ill people, and

---

[1] Ibid., p. 99.

usually overcome by our moral skill. If an evil possible
that the hard circumstances may come by pain to highness to
prove fit at or not solemn moving be expected and it be
a mistake to [illegible] ourselves all morals to
good or [illegible].

this is what permits a final re-appearance of the slippery slope argument. Mentally ill people may be capable of exercising some moral skill, but this is far outweighed by the handicap they so unluckily suffer. The disquieting suggestion is that the mentally ill person's situation differs only in degree from that of anyone who acts badly. The sane person who commits crimes may well suffer from milder, but equally unlucky handicaps; and to be consistent we should mitigate our blame in proportion to the severity of their bad constitutive or circumstantial luck. An excess of this kind of bad luck may, like insanity, provide a general excuse from responsibility, by impairing the unlucky person's ability to act on or even form qualitative evaluations. (Bad consequential luck threatens responsibility on particular occasions, and will be discussed separately in the next section.) In this way, any untoward act may be partly attributable to blameless handicaps, which would be uncovered if we were prepared to delve deeply enough into the malefactor's psyche and past.

The best way to refute this particular version of the slippery slope is to show that there is a clearly statable difference between infants and insane people and, on the other hand, normal, responsible adults. This will prevent the assimilation of all criminals into the same, excusable category as insane people. The criterion of responsibility which I developed in Chapter III tried to articulate this

this is what Bernoulli's principle is all about.

very difference, especially in section 4, where I offered the criterion as an explication of our intuitions regarding excuses from responsibility. It is important to see that Dennett explicitly rules out such solutions to the slippery slope argument. When he tackles the problem most directly,[1] Dennett confesses that he can find no principle to justify excusing those malefactors we normally excuse, while holding others responsible. Consistent application of the rationale behind existing excuses would require universal exculpation.

> So we must get arbitrary again, and draw the line - exactly where is no more important in this case than it is in the case of setting a legal age for drinking or driving. We must set up some efficiently determinable threshold for legal competence, never for a moment supposing that there couldn't be intuitively persuasive "counterexamples" to whatever line we draw, but declaring in advance that such pleas will not be entertained. We mustn't look too closely at the particular micro-details of the accused's circumstances, but just try to establish (crudely and swiftly) that in general this agent is deterrable, even though he was not deterred on this occasion.[2]

Of course, Dennett is not advocating a capricious, random system of responsibility and punishment. He proposes a system of punishment justified on purely utilitarian grounds. He, no more than anyone else, would not favor the punishment of infants or insane people. The reason for this, however, is that such punishment would not

---

[1] "Diminished Responsibility and Specter of Creeping Exculpation," Ibid., Chapter V, pp. 156-65.

[2] Ibid., p. 162.

further the utilitarian goals of deterrence, prevention, and reform. In fact, punishing such people would be counter-productive, by weakening public respect for such an unfair legal system. Crucially, there is no moral right to be spared from punishment which applies only to those people whom we traditionally excuse. These people just happen to be the beneficiaries of a social policy designed to minimize lawbreaking.

What can be said about this purely utilitarian view of responsibility as identified with amenability to effective punishment? (If there is such a thing as "pure" moral responsibility, isolated from the question of punishment, it too is assessed by Dennett in utilitarian terms.[1]) A comprehensive discussion of such a view would itself require a dissertation; and besides, I intend deferring my support of a partly retributive view of punishment to the next and final chapter. At this stage, I will confine myself to two brief comments.

First, Dennett's injunction to not "look too closely at the particular micro-details" of the situation revives a suspicion that was raised above in Chapter III, section 5: he is asking us to base our attributions of responsibility on ignorance. The principled criterion of responsibility I offer in Chapter III, on the other hand, does take detailed

---

[1]Ibid., pp. 166-9.

further the Australian Navy or Air Force, Brigadier Bung region." In 1966, Sjarifuddin, who would be

account of the capacities of the accused, and is therefore not open to this objection.

Second, our intuitions support my view that the distinction we make between responsible and blameless people is not arbitrary: *contra* Dennett, we feel that infants, insane people, etc., *deserve* to be excused. It is true, though, that my criterion of responsibility in terms of the capacity to act on justifiable qualitative evaluations does not draw a clear line between responsible and excusable malefactors. As I indicated in my discussion of insanity as an excuse (Chapter III, section 4c), determinations of, e.g., whether a desire could have been resisted, even in the case of sane people, can be very hard to make. However, as I then claimed, this difficulty seems entirely appropriate in the light of the uncertainty of our intuitions in hard cases. Moreover, the absence of a clear boundary does not disprove the existence of a principled distinction.

Extreme cases of bad constitutive and circumstantial luck are therefore not fatal for our moral framework, and do not create a slippery slope. Our reflective moral judgments, which I tried to capture in my criterion of responsibility in Chapter III, already *do* compensate for the effects of extreme handicaps. Moreover, because these are extreme handicaps, they are *rare*, and most of us, in sharp contrast, do not suffer from them, and are thus

account of the capacities of the accused, and is therefore not open to this objection.

Second, [illegible]

eligible for full responsibility for our actions. As I pointed out in the beginning of Chapter I, it seems quite appropriate for defense attorneys to ask a court for leniency in view of a criminal's troubled background. Unlike Dennett's utilitarian view, my criterion of responsibility clearly justifies this practice, and the distinctions we make between blameworthy and excusable criminals.

4. Extreme Cases of Bad Consequential Luck

Unlike extreme cases of constitutive and circumstantial luck, it is not clear that our reflective moral judgments already compensate for the effects of bad consequential luck. Bad constitutive and circumstantial luck may impair our ability to deliberate and act, and they tend to provide a general excuse or mitigating factor for all of someone's misdeeds. Consequential luck, on the other hand, in no way interferes with our deliberations, since it strikes after our actions. Furthermore, it only infects particular actions, and not all of them. The traditional excuses of ignorance, compulsion, and insanity tend to be restricted to factors which handicap us before we act. Moreover, because consequential luck strikes after we have acted, we have less opportunity to compensate for its effects by the use of "moral skill." In view of these differences, it is not

surprising that the most graphic examples of moral luck discussed in the literature - e.g., Nagel's drunken driver case, and Williams' discussion of Gauguin - tend to be cases of bad consequential luck. To complete my attempt to solve the problem of moral luck, I need to deal with these extreme cases of bad consequential luck. Dennett may have shown that such cases, being atypical, do not affect our attributions of responsibility in everyday cases; but we still need to explicate and justify our moral judgments in these hard cases themselves.

The problem in the drunk drivers case is to justify the difference in our attitude towards, and our punishment of, them. It will be recalled that they are equally drunk, but that one of them recklessly kills a child at a pedestrian crossing. The other driver would certainly have hit and killed the same child, but returns home without hurting anyone. They seem to be equally guilty, with luck being the only difference between them. Why should we treat one more harshly than the other? The traditional excuses of ignorance, compulsion, and insanity will be of little use, since they (fail to) apply equally to both drivers. The same applies to my criterion: both are equally capable of acting on qualitative evaluations. As for Dennett's notion of exercising our skill to compensate for the influence of luck, both drives were conspicuously devoid of skill in their drunken driving.

Another suggestion of Dennett's might be used to justify our different attitudes to the two drivers: "luck averages out in the long run."[1] This is no doubt true for the majority of us, who suffer no more than our fair share of lucky breaks and unlucky setbacks. Doubtless the unfortunate driver would in the future enjoy compensatory good luck; but probably not enough to make up for his manslaughter conviction. We are concerned here with the Ziggy's of this world, who are constantly beset by bad luck; and the Inspector Clouseau's, whose incompetence always miraculously leads to good results. Averaging out of luck does not apply to them. Similarly to Dennett, Coyne suggests that there is a sense in which the inequities caused by luck are, after all, fair. We are all equally subject to the vagaries of luck.[2] Thus we all have an equal chance to be caught up in a tornado, or struck by lightning. This is dramatically illustrated when a house is levelled by a tornado, while the neighboring one is left untouched. Yet we would hardly say to the person whose house has been destroyed, "I don't want to hear a word of complaint from you! Everyone was in danger, and someone had to lose!" We would, of course, extend every sympathy to the victim, precisely because she was the unlucky one. The problem created by moral luck (i.e., luck

---

[1] Ibid., p. 95.

[2] Coyne, op. cit., p. 8.

which influences moral assessment of our actions) is that we don't compensate for the relative misfortune of, e.g., the drunken driver who kills the child. Despite the fact that less luck is all that distinguishes him from the other driver, we do blame him more.

We might try to escape the charge of inconsistency by turning to an argument of Kant. Encountering bad luck gives one an opportunity to build strength of character,[1] and the unfortunate driver has an advantage in this respect over the lucky one. This is reminiscent of a familiar solution to the problem of evil: misfortunes and setbacks give people the opportunity to rise to the occasion, and they further God's goal of "soul making." However, even John Hick, in his defense of theodicy (i.e., the view that the existence of evil is compatible with rational belief in God), admits that setbacks

> sometimes . . . lead, instead, to resentment, fear, grasping selfishness, and disintegration of character. Therefore, it would seem that any divine purpose of soul-making . . . must continue beyond this life if it is ever to achieve more than a very partial and fragmentary success.[2]

Such appeals to the afterlife are beyond the scope of this dissertation. Indeed, it should not be surprising that Kant's argument should be unsatisfactory, since both Nagel

---

[1] Quoted by Nagel, op. cit., p. 33, footnote 8.

[2] John Hick, Philosophy of Religion (Prentice-Hall, 1963), p. 47.

and Williams explicitly cite Kantian views about morality as the source of the problem of moral luck.

A more promising approach to problem cases of moral luck makes use of the insight in the previous sentence. It appears in Judith Andre's highly original article.[1] Both Nagel and Williams have found an apparent inconsistency within our moral judgments in problematic cases. On the one hand, we feel that the lucky drunken driver is just as blameworthy as the other one; and that Gauguin's decision to leave his family can be assessed as reasonable or not relative to the evidence available to him at the time, regardless of its actual outcome. At the same time, we do blame the drunken driver who kills more than we blame the other one; and our final judgment on Gauguin will be influenced by whether he eventually succeeds as an artist. Nagel and Williams find the combination of these two pairs of judgments to be paradoxical, argues Andre, because they consider the first, Kantian pair to be more fundamental. From this Kantian perspective, the latter pair of judgments is indeed unfair; but these judgments are firmly backed by our moral intuitions - hence the antinomy.

Andre's novel proposal is that we should recognize two different, non-competetive strands within our moral judgments. Those of our judgments which do not compensate for the effects of moral luck - e.g., our belief that the

---

[1] Andre, op. cit.

driver who killed the child is more blameworthy than the other one - are not defective Kantian judgments. Rather they are made according to a different standard, which Andre calls "Aristotelian." Hence the two pairs of judgments considered above are not in conflict. Aristotelian judgments differ from Kantian ones in praising excellence of action and character, regardless of whether they are within the agent's control. The common thread in both kinds of judgment is that

> [moral] refers to an excellence of character such that the moral person is praiseworthy and emulable. He or she is a model for our children, and the kind of person which we would like our communities to foster.[1]

However, someone with a moral defect in the Aristotelian, as opposed to the Kantian, sense, is not necessarily deserving of blame or punishment. This accounts rather nicely for our judgments concerning constitutive luck. It is our Kantian intuitions that make us uneasy in praising or blaming people for character traits which are very likely beyond their control. At the same time, we do praise people for their good nature, knowing full well that this nature was partly shaped by constitutive (and also probably circumstantial) luck. These latter judgments, being based on Aristotelian grounds, do not conflict with the other, Kantian judgments.

---

[1]Ibid., p. 204.

driver who killed the child in auto accident must pay

The sense in which we find the two drunken drivers to be equally blameworthy is Kantian. We can now see that it is in a non-competitive, Aristotelian sense that we view one driver more severely than the other. Responsibility in Andre's Aristotelian sense does not call for punishment, as an expression of our disapproval. It merely creates an "obligation to rectify bad consequences."[1] This corresponds to Feinberg's notion of a penalty, as opposed to a punishment.[2] If no adequate compensation is possible - as is the case for the unfortunate drunken driver - a person is left with a "prosaic" sense of responsibility: a "feeling of inadequacy," and "a sense of diminished worth."[3]

I have three criticisms of Andre's argument. First, it is curiously weak in the very example about the drunken drivers which she discusses. On her analysis, the only difference in the responsibility of the two drivers is the "prosaic" sense which makes the unfortunate driver feel sad. In a Kantian sense, which determines blame and punishment, they are equally guilty. Yet we do blame and punish the unfortunate driver more heavily, which is precisely why this is a problematic case of moral luck.

---

[1] Ibid., p. 205.

[2] Joel Feinberg, "The Expressive Function of Punishment," op. cit.

[3] Andre, op. cit., p. 206.

The "Aristotelian" assessment is described by Andre as irrelevant to the issue of punishment. It gives rise only to a "prosaic" sense of responsibility, and so cannot justify our punishment of the drunken driver. Andre will surely have to fall back on her alternative explanation of the case. She suggests that there really is no moral difference between the two drivers. After all, a relative of the dead child will doubtless feel equally resentful towards all drunken drivers, recognizing that any of them could easily have been the one who killed their child.[1] This alternative explanation makes no use of her notion of an Aristotelian element to our moral judgments.

Second, it is not clear how the "Aristotelian" strand of our moral judgments differs from straightforward utilitarianism. The operative feature, for Andre's purposes, of the Aristotelian approach is that it permits judgments based on the actual consequences of actions, making no allowance for the influence of moral luck. Thus, for example, we criticize the unfortunate driver more heavily than the lucky one, despite the fact that only luck distinguishes them. In this respect, Aristotelianism is a

---

[1] Ibid., pp. 203-4.

consequentialist, utilitarian standard of appraisal.[1] If so, Andre's move of positing two separate, compatible elements within our moral judgments loses its plausibility. Kantian and utilitarian standards are certainly different, but they are usually viewed as rival approaches to ethics, and the burden of proof surely falls on those who argue that they are compatible. To escape this criticism, Andre must show how "Aristotelianism" contains non-utilitarian elements which are essential to it. Her notion that those who cannot compensate for causing harm should feel "a sense of diminished worth," for example, does not seem to be a utilitarian notion; but is it essential to Aristotelianism?

Third, there is a suspicion arising from the rather vague nature of Aristotelianism. Is the notion of an independent Aristotelian element of our moral judgments tailored in order to formulate an ad hoc solution to the problem of moral luck? Any pair of apparently contradictory beliefs can be reconciled by suggesting that they should be appraised by different standards. For such

---

[1]This is an over-simplification of utilitarianism. A utilitarian could appeal to the difference between appraisals of actions, and appraisals of agents. (See John Stuart Mill, Utilitarianism (Hackett edition, 1979), p. 18, footnote 2.) For example, suppose that someone, despite exercising great caution, unluckily causes an accident. Although the outcome is of course bad, it would be prudent not to criticize him as an agent, since he exercised the very caution we want to encourage. However, such a reply is not available for our drunken driver, who was reckless.

reconciliations to be effective, both standards of appraisal must be clearly articulated, and plausible. While Kantian standards meet this requirement, Andre needs to be clearer on exactly what Aristotelian standards are.

5. A Partial Solution to the "Drunken Drivers" Case

There is a much simpler, more plausible justification of most of our attitudes in the drunken drivers case, which will help to re-focus our examination of consequential luck. This account makes no attempt to identify different standards of evaluation within the same moral judgment (e.g., Kantian vs. Aristotelian). Instead, it gives a single standard of responsibility, consisting of three conditions. This account is not original, but its application here will be invaluable. It is Feinberg's "Tri-Conditional Analysis" of when a harm can be attributed to an agent as "his fault."[1] A person A is responsible for a harm X if and only if:

1. A is at fault for creating the risk or certainty of harm (The Fault Condition).
2. A caused X (The Causal Condition).

---

[1]This is the main focus of Feinberg's "Sua Culpa," op. cit. For an initial exposition of the Tri-Conditional Analysis, modifications of the conditions, and discussion of how to apply them, see pp. 195-207.

3. X was within the risk for creating which A is at fault, and A caused X by virtue of this fault (The Causal Relevance Condition).

We can safely ignore any complications which would require modification of these conditions, since we are dealing with a relatively simple case.

The driver who killed the child plainly meets all three conditions. The very reason we abhor drunken drivers is that they recklessly create the risk of precisely what did happen: mutilation and death. By hypothesis, it was because of his recklessness that death resulted. Had he been sober, his reaction time would have been quicker, he would have seen the red light and the child would have lived. His fulfillment of these three conditions accurately reflects our intuitions that "he got what he deserved."

What distinguishes him from the lucky driver is not that he is responsible in some vague, prosaic Aristotelian sense. It is the simple fact that he meets the causal condition by having caused the child's death. The causal condition can itself be justified (though it need not be) on retributive, Kantian grounds. A necessary condition of retribution must surely be that a harm actually be caused. While attempted murder is itself punishable, the greater harm caused by murder is at least part of the reason why,

on retributive grounds, murder deserves a heavier punishment.

It will be objected that I am ignoring the main point of the case: the driver is unlucky to have killed the child. My answer is simply that he is in no way unlucky. The driver who misses causing any harm is certainly lucky, and this may tempt us to view the killer of the child, by contrast, as unlucky. However, in view of his actions, the latter cannot plausibly claim to be the victim of bad fortune. The Tri-Conditional Analysis gives a particularly easy method for determining whether a person is lucky or unlucky for a given action. A person is lucky if the harm, if any, he causes is considerably less than the harm for creating the risk of which he is at fault. A person is unlucky if the harm he causes is beyond the risk of harm for creating which he is at fault. The driver who kills the child is neither lucky nor unlucky. The harm which he causes is precisely the harm of which he is guilty of running the risk.

The beauty of the Tri-Conditional Analysis is that the influence of bad consequential luck on moral responsibility is systematically eliminated by the causal relevance condition. Contrary to Nagel's, Williams', and Andre's fears, we can and do compensate for any unfair influences of bad consequential luck on our judgments of responsibility. Williams may be correct that Gauguin's

punishment.

retrospective feelings will depend on fortuitous factors beyond his control; but using the Tri-Conditional Analysis we can make a fair Kantian judgment as to his responsibility, in view of the risk he took. In the first part of the next section, however, I will argue that the influence of bad moral luck should not be altogether eliminated from our moral framework.

Still, it might be insisted, I am ignoring the main problem: and this is the blatant injustice of the lucky drunken driver getting off Scot free, or at least more lightly than the less fortunate driver. The lucky driver was just as guilty of recklessness, and only escaped the same fate by good luck. While we readily accept the punishment of the driver who killed the child, when we turn to a comparison with the lucky driver, our intuitions are offended. This is the problem of good moral luck, and the question of whether we should prevent it from getting people like the lucky driver off the hook. I will discuss this in the second part of the next section.

6. Morality Without Luck

The final stage in my attempt to gradually defuse the problem presented by various forms of moral luck consists in a more negative approach. Until now, I have tried to justify or compensate for the role played by luck in our moral judgments. Now I will point

out two pernicious consequences of purging our whole moral framework of the influence of luck.

a. Bad Luck, and Personal Causal Connections

I argued in section 3 that very bad constitutive and circumstantial luck can be compensated for by my own criterion of responsibility. To the degree that a bad genetic "hand," a poor environment, or repeated setbacks throughout one's life reduce one's capacity for acting on qualitative evaluations, we diminish our blame for his misdeeds. Bad consequential luck, I argued in section 5, is eliminated from our moral assessments of actions by the Causal Relevance Condition of the Tri-Conditional Analysis. Here I will draw on a theme in Williams' work, and argue that there is an area of our moral framework where the influence of bad moral luck should remain. This area is our attitude towards our own actions.

Williams gives an insightful analysis of the retrospective attitudes of regret and agent-regret which we may adopt towards the harms we cause.[1] Regret is the impersonal feeling of "how much better if it had been otherwise,"[2] which we experience when viewing other

---

[1] Williams, op. cit., pp. 27-31.

[2] Ibid., p. 27.

people's misdeeds, and also some of our own involuntary misdeeds (e.g., crimes committed under hypnosis: "It wasn't me who did it.") Agent-regret is confined to our own misdeeds, usually those that we commit voluntarily. However, Williams' surprising contention is that we may also feel agent-regret for actions that are not fully voluntary. These may include accidents, and actions of which we are not fully in control, due to ignorance or various kinds of compulsion.

For example, Williams considers the truck driver who blamelessly kills a child who suddenly runs in front of the vehicle. Despite our sympathy for the driver,

> there is something special about his relation to the happening, something which cannot be eliminated by the consideration that it was not his fault.[1]

Despite his lack of control over the outcome, the action is his. This is part of Williams' main thesis, described above in section 1 of this chapter, that our agency includes the unintentional and fortuitous consequences of our actions. (Williams also discusses exactly when it is appropriate to view our actions impersonally, with simple regret. We can safely ignore such detail here.)

The particular feature of Williams' thesis that I want to bring out is his suggestion that it is healthy to experience agent-regret for some of one's less than fully

---

[1] Ibid., p. 28.

voluntary actions, and their consequences. From a Kantian viewpoint, the truck driver is blameless, and is justified in viewing his actions externally, with regret only. Still, we would be suspicious if he experienced no twinges of agent-regret, or was "too blandly or readily moved" by our reassurances that it wasn't his fault, and that he shouldn't take it personally. Someone who viewed her actions in purely Kantian terms would consider only her intentions, and would suffer no agent-regret for unlucky consequences of her actions. Williams is suggesting that such a person would be morally insensitive. This is not the moral insensitivity displayed by a psychopath, who can barely understand moral concepts. It is rather the failure to experience certain emotions which we expect morally decent people to be able to feel. The importance of emotions to our moral judgments and to our moral worth is an important theme for Williams.[1]

The idea that we sometimes ought to feel agent-regret, even though we are morally blameless, also appears in others of Williams' writings. This is doubtless one aspect of what Coyne was getting at when she referred to his theme of "the reality of tragedy in human life."[2] In "Ethical

---

[1] See especially "Morality and the Emotions," Problems of the Self (Cambridge, 1973).

[2] Coyne, op. cit., p. 5.

Consistency,"[1] he considers moral dilemmas in which weighty moral considerations pull us in opposite directions. His main point is that, even though a close examination may show one consideration to be clearly more weighty, there will be a "moral cost" involved in violating the other moral obligation. This other obligation does not simply disappear when it is outweighed; rather, it remains as an unsatisfied moral "residue." In such situations, it is quite appropriate for the agent to feel agent-regret, even though he has acted in the morally optimal way. That he found himself in such a moral dilemma is an instance of bad circumstantial luck. Again, on pain of moral insensitivity, his attitude towards his own action should not discount the contribution of (bad) moral luck.

The same insight can be applied to the Gauguin example, if it should turn out that he has bad consequential luck. His decision and action may have been impeccable from a Kantian viewpoint: he may have correctly weighed the probabilities and values of different outcomes for his future success as an artist. However, if bad consequential luck (as defined in section 5) intervenes, he ought to feel agent-regret. The same analysis can be applied to all victims of bad consequential luck, however blameless they may be.

---

[1] Problems of the Self

An objection which I made to Andre re-emerges here: how does the source of this agent-regret differ from the purely utilitarian judgment that we are to blame, or at least to feel bad, for causing bad outcomes? That agent-regret does differ from utilitarian concerns is illustrated by the related notion of integrity, which Williams uses as an argument against utilitarianism. He discusses two examples of moral dilemmas in which the utilitarian solution is clear.[1] Both cases involve an agent preventing a catastrophe by himself deliberately causing a smaller, but significant, amount of harm. Since it is not obvious that the utilitarian solution is the correct one - it would not be outrageous for the person involved to refuse to perform a deliberately harmful act - Williams concludes that utilitarianism does not reflect the complexity of our moral decisions. The element it omits is integrity: the special relation we have to our own actions, as opposed to bad outcomes caused by other people.[2] That one finds oneself in an

---

[1] J.J.C. Smart and Bernard Williams, Utilitarianism: For and Against, (Cambridge, 1973, pp. 97-100).

[2] The notion of integrity is open to the charge that those who are very much concerned with it are guilty of "moral self-indulgence." To refuse to prevent a catastrophe by causing a small amount of harm can be criticized as a selfish concern with the purity of one's own "moral slate." Williams articulates and replies to this criticism in "Utilitarianism and Moral Self-Indulgence," Moral Luck.

integrity-threatening dilemma is another instance of bad circumstantial luck.

I suggest that integrity is of a piece with agent-regret. What they have in common is what I call personal causal connection: the special concern we have for what we actually do and its consequences, regardless of extenuating circumstances. This concern cuts across both Kantian and utilitarian appraisals. An action may be impeccable from a Kantian viewpoint - for example, Gauguin's decision, assuming that it was carefully and honestly made - and yet still give rise to agent-regret. Similarly, the happiness-maximizing action endorsed by utilitarianism may still violate the integrity of the agent. Both instances of personal causal connection are subject to (bad) luck: consequential luck in the outcome of one's actions (agent-regret); and circumstantial luck in facing integrity-threatening situations. We now have a clearer idea of what would be lost by purging our subjective attitude to our own actions of the influence of bad moral luck. Doing so would fail to do justice to the importance of personal causal connections in our moral framework, and would be morally insensitive.

So bad moral luck should indeed be compensated for in our moral judgments. Its influence should remain, however, in our attitude towards our own actions, in the form of our concern with personal causal connections. This is, I

believe, the proper place for Andre's notion of a non-Kantian element in our moral framework. The notion of personal causal connection, better than Andre's "Aristotelianism," fills the need for a non-Kantian, non-utilitarian strand of moral appraisal.

b. Responsibility without Causation

I turn finally to the problem of good moral luck, which arose in my discussion of the drunken drivers example in section 5. To begin with, good constitutive and circumstantial luck present no problems. While we should and do compensate for the effects of bad luck of these kinds on our moral assessments (see section 3), it would be pointless and mean spirited to try to eliminate the effects of good luck. A child inherits a genius-level I.Q. from his college professor parents. He is brought up in a safe, comfortable, caring environment. He is sent to the best private schools, and goes to a prestigious Ivy League college. He follows his parents' footsteps into the academic world, and a tenure-track position opens up for him at his alma mater just as he is leaving graduate school and entering the job market. Throughout his life he enjoys the best of health, and is blissfully happy with the witty, intelligent and sensitive woman he met and married in graduate school. This man has plainly been blessed with

the most extraordinarily good constitutive and circumstantial luck. He will be much less tempted to do bad things, and hence incur moral criticism and punishment, than poor, abused, unemployed, and otherwise disadvantaged people. He certainly does not deserve such good fortune, but, crucially, he does not deserve not to have it, either. No principles of justice are offended by people's enjoyment of such good luck; whereas it would be arbitrary and unjust, in the absence of overriding reasons, to try to strip people of the benefits of such good luck.[1] The ultimate consequence of an effective policy of equalizing the effect of this kind of good luck would be a society in which everyone is equal in talent, health, income, education, family happiness, love life, etc. - hardly a feasible or desirable target.

---

[1]However, such overriding reasons may sometimes be available. In the name of equality of opportunity, it may be justified to restrict the degree to which a millionaire should be allowed to give his daughter a "head start" in life. Nepotism in the job market is rightly frowned upon; and it would not be outrageous (though hardly practicable in today's U.S.), to advocate the abolition of private schools for children up to the age of, say, ten. It could be argued that every child has a right to begin on an equal footing, and that expensive private schools give the children of the rich an unfair advantage. Such policies could doubtless be viewed as the attempt to strip people of the benefits of good moral luck. However, my point is that denying people the benefits of good moral luck would not be an end in itself; rather, it would be a means of achieving certain social goals (e.g., equality of opportunity). My objection is against neutralizing good moral luck for its own sake.

Good luck appears to be pernicious only in the case of consequential luck, and most often not even in this case. Winning a lottery is an incredibly lucky consequence of buying a ticket. Becoming a millionaire is a most unlikely, fortunate consequence of starting a chain of pizza stores. However, for the same reason as stated in the previous paragraph, we would not dream of trying to eliminate the effects of such luck, even though the enjoyment of it will very likely make it much easier for these people than for most to be good citizens.

Good luck becomes troublesome in the case of people like the lucky drunken driver who create the risk of great harm, which they avoid only by fortunate circumstances. In terms of the Tri-Conditional Analysis, such people satisfy the fault condition, but escape responsibility by luckily failing to meet the causal condition. It seems unjust that the lucky driver receives at most a drunken driving conviction, while the other driver, who is no more at fault, is charged with involuntary manslaughter. Our desire not to let people like the lucky drunken driver off the hook suggests a drastic modification of the Tri-Conditional Analysis. Why not drop the causal condition (and ipso facto the causal relevance condition), and base responsibility solely upon fault? This would allow us to punish the lucky driver as heavily as the other one. I will call the motivation for eliminating good moral

luck from our attributions of responsibility for harm "the retributive principle" (following Feinberg).[1] Such an elimination is undesirable for three reasons.

First, let us re-consider Williams' truck driver, who blamelessly kills a child when she suddenly runs in front of the vehicle. With respect to the child, the driver is in no way at fault, and so the retributive principle seems to justify our intuition that the driver should not be punished for her death. But what if the driver was at fault in other respects? For example, he may have been knowingly driving with severely defective brakes, which could have failed at any time. (For simplicity's sake, let us assume that the brakes did not fail when he hit the child; there simply was not enough time to stop.) It was only very good luck that prevented this fault, and maybe others beside it, from resulting in extensive property damage, injury, or death. According to the retributive principle, his faulty enjoyment of this good luck creates a substantial "moral debt"; and a policy of "linkage" could be used to justify making him pay this debt by holding him responsible for the child's death. A variation which better illustrates the notion of "paying a moral debt" would be an accident in which the driver swerves to avoid a dog, and hits a parked car, causing extensive damage. If

---

[1] Feinberg considers "the retributive theory of torts," in which liability, but not necessarily blame, is based solely on fault. Feinberg, op. cit., pp. 212-21.

the damage he thus faultlessly causes is of roughly equal value to the damage which, despite his faulty driving, he luckily avoids causing, the policy of linkage would justify holding him liable for the damage to the car. Basing responsibility and liability solely on fault has even more counter-intuitive results. Why should we confine our attention to the driving faults of the person in the previous example? He may have been a perfect driver, but persistently acted recklessly in other areas of his life. He may, for example, have illicitly borrowed funds from the bank where he works, and used them for gambling. Fortunately, he has always won at the casinos, so no one has lost any money. Provided there is monetary equivalence between the financial damage he luckily avoided, and the damage to the car which he unluckily caused, the retributive principle seems to justify holding him liable for the latter. Stranger still, we could draw on his "moral debt" to hold him liable for the damage other people unluckily cause.

The source of these problems is the fact that the elimination of the causal condition removes the need for any connection, aside from monetary equivalence, between the agent's fault, and the harm for which he is held liable. This seems to violate another tenet of retributivism: the principle that people should only be

blamed and punished for what they actually do.[1] Retribution, viewed as the expression of our condemnation of a specific action,[2] seems to lose its point when linkage is used to punish us for unrelated harms. The initial motivation for dropping the causal condition was to base responsibility and liability, using the retributive principle, on fault alone. Ironically, the causal condition itself seems to be required by other tenets of retributivism.[3]

Second, there are doubts about the practicability of enforcing the retributive principle. A system of liability based on the retributive principle is indeed possible: Feinberg describes, somewhat in the spirit of a caricature, "the retributive theory of torts."[4] (Tort liability does not imply any condemnation of those held liable, and so may escape the previous criticism that the retributive

---

[1]This is what I call in Chapter V the "causal principle."

[2]See Chapter V for my account of "the expressive principle."

[3]In H.L.A. Hart's terminology, the motivation for basing responsibility solely on fault, by dropping the causal conditions, is "retribution in general justifying aim." The principle with which it conflicts in this case, namely that only those who actually commit crimes may be punished, is part of "retribution in distribution." The latter principle restricts whom we are justified, on retributive grounds, in punishing. See "Principles of Punishment," Punishment and Responsibility (Oxford, 1968), esp. pp. 8-13.

[4]Feinberg, op. cit., pp. 215-7.

blamed and punished for much of the destruction and

principle would violate other retributive tenets by misplacing moral condemnation.) Feinberg describes a more systematic, thorough version of the system of linkage which I illustrated above. At the end of each year, each person's "demerits" are totalled, and he is forced to pay a corresponding fine into the "community fund." This fund will be used to compensate all the previous year's victims of harm, regardless of who caused the harm. This system has at least some plausibility when applied to the cases we have considered of property damage, which can be assigned a definite monetary value. However the retributive theory of torts must take into account all moral faults, in order to determine our total moral debt for the year. How could we possibly assign a "demerit value" to faults such as telling lies, laziness, reckless driving, rudeness, etc.? And how much money out of the community fund is needed to compensate the parents of a child killed in a car crash? Furthermore, even if we could devise a scale of moral demerit, on which each fault would be assigned a certain number of points, merely keeping a record of people's moral faults would require 1984-style vigilance of our everyday actions.

Third, the retributive principle is excessively severe in the moral condemnation it requires us to make; or, in the case of tort liability, in the burdens of liability it directs us to place on people. Instead of allowing lucky

breaks to get us off the hook, it systematically tries to neutralize these breaks. Plea-bargaining or executive clemency, for example, would have no place under it. Indeed, forgiveness in general could be viewed as good moral luck for its recipient, and would hence be excluded by the retributive principle. A famous application of this principle occurs in Kant's argument for capital punishment. Kant insists that, even if society were somehow to become dissolved, we must execute all murderers left in prison "so that everyone will duly receive what his actions are worth."[1] Still, a determined retributivist (who also supported the death penalty) would likely see this as an advantage for the retributive principle, rather than as a reductio ad absurdum of it. However, there is a more extreme and less palatable consequence of regarding the punishment of every moral fault, even if it issues in no bad consequences, as an end in itself. As Feinberg points, it would follow that

> . . . a world in which everyone flourishes regardless of moral condition would be intrinsically inferior morally to a world in which all and only the good flourish and all and only the bad suffer. If everyone without exception is a miserable sinner, then it is intrinsically better that everybody suffer than that everybody, or even anybody, be happy.[2]

---

[1] Immanuel Kant, The Metaphysical Elements of Justice (1797), John Ladd translation, p. 102.

[2] Feinberg, op. cit., pp. 217-8.

The elimination of good moral luck would indeed require this drastic conclusion, which is surely too harsh. If sin leads to no bad consequences, i.e., no suffering or unhappiness, what moral purpose would be served by making sinful people unhappy? In his defense of hedonistic utilitarianism, J.J.C. Smart makes a similar point with his example of the "deluded sadist."[1] Sin is objectionable because of its universally bad consequences; but stripped of these bad consequences, Smart argues, sin is no longer a supreme evil. This final argument against the retributive principle, then, is that its demand for punishment of even harmless sinners serves no other moral purpose than vindictiveness. Hence the elimination of good moral luck, which is motivated only by the retributive principle, is not justified. While we would scarcely admire the faulty recipients of good moral luck whom we have considered, we cannot rightfully deny them the benefits of this luck.

7. Overview

I began this chapter with an account of the threat posed by moral luck to our moral framework (section 1). Dennett's solution, which consists in emphasizing our ability to compensate for bad luck by using moral skill, successfully heads off the main threat posed by moral

---

[1] Smart and Williams, op. cit., pp. 25-6.

luck: not all of our actions are at the mercy of luck (section 2). However, an extension of traditional excuses from responsibility, especially insanity, threatens a re-appearance of the slippery slope: all misdeeds may be attributable to bad constitutive and circumstantial luck. While Dennett's utilitarian approach does not give a satisfactory answer to the slippery slope, the criterion of responsibility I presented in Chapter III halts the slippery slope. It explains and justifies a clear difference between blameworthy and excusable malefactors (section 3). Still, extreme cases of consequential luck present problems which are addressed neither by Dennett's general solution to the moral luck issue, nor by my criterion of responsibility. Moreover, they are not handled satisfactorily by any existing analysis (section 4). A simple, effective treatment of the problem posed by bad consequential luck is provided by an application of Feinberg's "Tri-Conditional Analysis" (section 5). Finally, there are two reasons for not completely eliminating the influence of luck on our moral judgments. First, it would be morally insensitive to fail to appreciate the importance of our "personal causal connection," with harms, even if they are the blameless product of bad luck (section 6a). Second, the elimination of the effect of any kind of good luck on our moral

judgments would be purely vindictive, and unjustified as an end in itself (section 6b).

This chapter completes the main part of my defense of compatibilism. In Chapter II, I explained and endorsed Dennett's naturalistic account of human agency. I argued that he successfully showed that the conditions necessary for moral responsibility are perfectly compatible with universal causation. In Chapter III, I expanded on, and at the same time began to diverge from, Dennett's account. I offered a more detailed account of the structure of our desires and will in the process of self-determination. Out of this account I produced a (defeasible) sufficient condition for responsibility. This goes beyond Dennett, in providing a framework for deciding whether specific people are morally responsible for particular actions. However, this condition of responsibility permitted a re-emergence of the hard determinist argument, which also applies to Dennett: is not the ability to exercise self-determination itself subject to causes beyond our control? Dennett's ultimate solution to this challenge is pragmatic, and this signals my greatest disagreement with him. Whereas he is content to argue for the rationality of acting as if we are in control of our destinies, I question the fairness of holding people responsible for actions which are, ultimately, the product of luck. In the present chapter, I have offered a comprehensive, non-pragmatic solution to

this problem of moral luck. It is not true that all of our actions are subject to luck. Furthermore, we already compensate for the effect of bad moral luck on our moral judgments, while the contribution of good moral luck is generally perfectly fair. My treatment of moral luck constitutes my final reply to the hard determinist challenge.

In the course of my discussion of moral luck, my disagreement with Dennett's utilitarian conception of responsibility and punishment has repeatedly surfaced. In the next and final chapter, I will draw together these criticisms; and then, more importantly, I will argue that the retributive view of punishment which I advocate in its place is consistent with my whole compatibilist approach.

this problem of moral luck. It is [illegible]

actions [illegible] good or wrong. [illegible]

con[illegible] [illegible]

[illegible]

CHAPTER V. COMPATIBILISM AND RETRIBUTION

To complete my dissertation, I will diverge from most compatibilist accounts of the determinism/moral responsibility problem, including that of Dennett, by advocating a partly retributive view of punishment. (However, this does not affect my endorsement of most of Dennett's naturalistic solution to the moral responsibility problem.) I will explain why most compatibilists have, erroneously in my view, felt constrained to adopt a purely utilitarian view of punishment. Finally, I will show that, on the contrary, my version of compatibilism is quite consistent with retributive justifications of punishment.

1. Utilitarian and Retributive Justifications of Punishment

A first approximation to the distinction between the two approaches to punishment is that the utilitarian view is forward, and the retributive view backward-looking.

A utilitarian is concerned with how, in view of the crime, to improve future life for the criminal, victim, and society as a whole. Punishment is justified by its beneficial consequences: deterrence of the criminal and other potential wrongdoers from future crimes, and hopefully reform of the criminal. If punishment would not

further any of these goals, however, then it is not justified, no matter how heinous the crime for which it is being considered. The retributive notion of "paying back" the criminal for his misdeeds is criticized as a primitive desire for revenge, albeit expressed in more acceptable, civilized terms. Similarly, the doctrine of "an eye for an eye," which is held by some retributivists (including Kant), is criticized as barbaric. (In the next section I will show how this criticism can be avoided.) On a utilitarian view, a major goal is to help the criminal to resume a useful and harmless life. The preoccupation of retributivists with evening the score for past crimes is viewed by utilitarians as harsh and unsympathetic.

Retributivism is not as easy to characterize. It regards punishment as an end in itself, which does not require justification by any good consequences it may have. It is backward-looking in that it aims at giving the criminal what he deserves for his misdeed, and at restoring fairness and justice. Here are six principles associated with a retributive view, not all of which need be accepted by any given retributivist.[1]

---

[1] See Stanley I. Benn's article "Punishment," in Paul Edwards (ed.), The Encyclopedia of Philosophy (Macmillan, 1967), Vol. 7, pp. 29-35, for more complete accounts of the rival theories of punishment. This article also contains useful detail on the retributive criticisms of utilitarianism considered in the next section.

1. Criminals deserve to be punished. (The Desert Principle)
2. Punishment of criminals is owed, as a matter of fairness, to law-abiding citizens who have made sacrifices in obeying the law. (The Justice Principle)
3. Punishment expresses society's moral condemnation of the criminal. (The Expressive Principle)
4. Punishment may only be applied to people who have actually committed crimes. (The Causal Principle)
5. Punishment may only be applied if the criminal was in full control of his actions, and capable of obeying the law. (The Capacity Principle)
6. Punishment should be graded to correspond to the severity of the crime. (The Proportionality Principle)

2. A Critique of a Utilitarian View of Punishment

A detailed critique of utilitarianism would itself be a topic for a complete dissertation. My aim in this section is to indicate, without extensive discussion, why I favor a view of punishment which is at least in part retributive. I will explain what I take to be the most telling criticisms of utilitarianism in the literature, and add two original

1. Criminals Insane, Ex Parte, [illegible]. [illegible] Report
[illegible].)
2. Punishment of [illegible] ... [illegible]

criticisms which have emerged from my defense of compatibilism.

The most familiar criticisms of a utilitarian view of punishment involve arguing that it fails to do justice to intuitively plausible tenets of retributivism, such as the ones listed above. Thus, for example, it is claimed that, in extreme circumstances, a consistent utilitarian would have to approve of the punishment of the innocent (violating the causal principle). Similarly, the punishment of clearly incompetent people who cannot reasonably be expected to obey the law may be required for the purpose of deterrence (violating the capacity principle). Equally implausibly, a purely forward-looking view may call for, e.g., punishing chronic tax evaders more heavily than murderers whom we are sure will never repeat their crime (violating the proportionality principle).

Such criticisms, which are usually based on extreme circumstances unlikely to arise, are weak as they stand. The standard utilitarian reply is to show how these offensive consequences can be ruled out on purely utilitarian grounds. Thus punishing the innocent or the incompetent, and excessive punishments for trivial offenses, can be opposed by a utilitarian on the ground that they are likely to be counter-productive by causing public fear and resentment, and lack of respect for the law. Alternatively, a consistent utilitarian could bite

[illegible] which [illegible] by [illegible]

comparisons.

The [illegible]

[illegible]

the bullet and admit that, in the very unlikely event that such extreme circumstances ever arose, the counter-intuitive utilitarian policy would indeed be justified.[1]

However, a deeper, more telling criticism emerges from the first utilitarian reply to these objections. This solution to the unpalatable practices endorsed by a simplistic utilitarian theory is contingent upon these practices having bad overall consequences. This reply surely misses the force of the counter-examples. Such practices are wrong simply because they are unfair and unjust, regardless of whether they can be ruled out by a more reflective version of utilitarianism.[2] The perception that these practices would be unfair and unjust is based upon everyday intuitions, and hence does not beg the question by assuming the truth of retributivism. The second utilitarian response, which consisted in biting the bullet, is also implausible to the extent that it violates these everyday intuitions.

---

[1]The latter alternative is the line taken by J.J.C. Smart in reply to the objection that utilitarianism would justify punishing the innocent. See Smart and Williams, Utilitarianism: For and Against (Cambridge, 1973), pp. 67-73.

[2]This point is made by Benn, op. cit., p. 31; and by H.L.A. Hart, even though he goes on to defend what he considers to be a broadly utilitarian view, in "Murder and the Principles of Punishment," Punishment and Responsibility (Oxford, 1968), pp. 75-9.

Moreover, it needs to be considered why, as utilitarians claim, the practice of punishing the innocent and the incompetent would lead to public unrest and outrage. It is because, quite simply, the public would view such practices as unfair. Dennett is certainly correct when, endorsing the utilitarian approach I am criticizing, he points out that "The public perception of the fairness of the law is a critical factor in its effectiveness."[1] But it is surely taking a cockeyed, mistaken view of the situation to argue that the fairness of the law is only important as a means to its effectiveness. The utilitarian goal of public acceptance of the law is dependent on the public's prior acceptance of the very retributive principles which utilitarians claim are not needed to justify our system of punishment.

H.L.A. Hart has argued that these and similar criticisms are not fatal to a utilitarian view of punishment.[2] His innovative suggestion is to distinguish between the "general justifying aim" of punishment, and, on the other hand, its "distribution." He would concede that the preceding criticisms prove the inadequacy of

---

[1] Elbow Room, p. 163.

[2] This is a central theme which runs throughout Punishment and Responsibility. See especially "Prolegomenon to the Principles of Punishment," "Murder and the Principles of Punishment," pp. 71-83, "Punishment and the Elimination of Responsibility," and "Postscript: Responsibility and Retribution," pp. 230-7.

utilitarianism in determining the distribution of punishment, i.e., whom, and in what circumstances, we may justifiably punish. He would certainly accept the causal and capacity principles as restricting the justifiable use of punishment. He insists, however, that this is perfectly compatible with holding a utilitarian view on the "general justifying aim" of punishment. The whole institution of punishment, he argues, is best justified by the utilitarian goals of protecting society and reforming the criminal. Retributive principles are needed, though, to justify particular cases of punishment. Hart is particularly concerned with what I have called the capacity principle, which he calls the mens rea requirement: only those who act voluntarily and with the required "mental element" may be punished. No amount of social defense would justify the violation of this principle, except for rare strict liability offenses.

However, despite the ingenuity of Hart's distinction between the general justifying aim, and the distribution, of punishment, it does not save utilitarianism from criticism. Even utilitarianism as a general justifying aim of punishment is open to two objections which have appeared in the literature. The second objection also serves as a reply to an objection which Hart has himself levelled at retributive justifications of punishment.

Herbert Morris has argued persuasively that a purely forward-looking justification of punishment would fail to treat criminals as persons.[1] Such a justification would have us view criminals not as malefactors deserving punishment, but as sick beings in need of treatment. People whom we currently regard as rational, autonomous wrongdoers would be assimilated to the insane, children, or worse still, mere objects. Ironically, punishing and blaming criminals for their misdeeds is a sign of respect for their autonomy, which is violated if we treat them in the same way that we currently treat those whom we do not hold responsible for their crimes. Hart himself points out the undesirability of this assimilation of rational wrongdoers to sick people, and argues that the retention of mens rea requirements (which are based on the capacity principle) is necessary to avoid it.[2] However, he insists that a forward-looking utilitarian general justification of punishment, provided that it is supplemented with the mens rea requirement, does indeed treat criminals as persons.

It is not clear that the inclusion of the mens rea requirement is sufficient to avoid Morris' criticism of a

---

[1]Herbert Morris, "Persons and Punishment," The Monist, Oct. 1968, reprinted in Richard Wasserstrom (ed.), Today's Moral Problems, 2nd edition (Macmillan, 1975).

[2]"Punishment and the Elimination of Responsibility, op. cit., pp. 182-5.

utilitarian general justification of punishment. Taking into account thoughts and intentions will certainly avoid the blatant insult of treating criminals "merely as alterable, predictable, curable or manipulable things" (my emphasis).[1] However, the mere consideration of our intentions does not guarantee that we are, unlike insane people and children, granted the respect which is due to us as rational, autonomous adults, i.e., as persons. In my discussion of Dennett (Chapter II, section A2), I pointed out that a psychologist adopts the design stance in describing a psychotic person's behavior. Such explanations do indeed refer to the (unconscious) intentions of the patient, but without attributing to him any rationality or coherent purpose. Similarly, it will be possible for those in charge of dealing with criminals in a purely utilitarian system of treatment to consider their intentions, but only as symptoms of pathology, of which they must be cured. An ominous possibility in such a system, as Morris points out, is that the motives of lawbreakers who claim to have a reasoned justification for their actions will be regarded in the same light as those of any criminal. Instead of taking the justifications which, for example, political protestors offer for their crimes at face value (e.g., an anti-nuclear demonstrator's desire to reduce the threat of annihilation), judges would

---

[1] Ibid., p. 183.

militarian anxiety Identification of "Dunkmoor." Capt.
into account examples. The increasing city celebrity
the pigeon stance of "Legal" citizens. "Public
Ital. Prof. Pro. Journ. de Droit ... "Dunkmoor."

determine how best to "cure" them of their anti-social beliefs. Even their very arguments to support their political protests, and their claims that they do not require "treatment," will be viewed as pathological symptoms.[1] This is reminiscent of the way that political dissidents are allegedly treated in Soviet "psychiatric hospitals."

The only way to avoid this dangerous and condescending form of paternalism is to take the protestor's arguments seriously, hold him responsible, and blame him for his actions. This requires the acceptance of at least two principles which belong to retribution as a general justifying aim of punishment. First, criminals deserve to be punished (the desert principle). Morris, following Kant and Hegel, defends its paradoxical corollary, the "right to be punished."[2] Second, the punishment must express our disapproval of the crime (the expressive principle). Blaming him shows that we regard his action as autonomous, rather than as an excusable symptom of disease. Accepting these retributive principles would be to depart from a utilitarian view of the general justifying aim of punishment.

---

[1]Morris, in Wasserstrom, op. cit., pp. 480-1.

[2]Ibid., pp. 472-82.

determine pos pert to "core" tmos of trait continuity
pelfide. Even [illegible] on [illegible]
political [illegible]
region [illegible]
[illegible]
[illegible]

The need for the expressive principle has been defended on different grounds by Joel Feinberg.[1] Though Feinberg does not present it as such, it can be viewed as an indirect attack on a utilitarian view of punishment, which has no room for this principle.

To view punishment, as would utilitarians, merely as the infliction of hardship in order to reduce future crime, argues Feinberg, ignores a crucial distinction between penalties and punishment. A penalty is the infliction of hardship without any disapprobation: e.g., parking tickets, fouls in various sports, strict liability offenses whose purpose is regulative. A punishment differs in that it has in addition a "symbolic significance": it expresses our condemnation and disapproval of the forbidden action. This denunciatory element certainly adds to the utilitarian goal of deterrence: as Mill has famously argued, moral sanctions can be even more coercive than overt legal ones.[2] In this sense, then, utilitarianism allows for the expressive function of punishment. However, any expression of disapproval symbolized by punishment is not an end in itself: it is of value, according to the utilitarian, only insofar as it increases social

---

[1] "The Expressive Function of Punishment," Doing and Deserving.

[2] John Stuart Mill, On Liberty, esp. Chapter I, "Introductory".

defense.[1] Yet its effectiveness as social defense depends on the public's prior acceptance of the expressive principle. In this way, utilitarian goals are subservient to this principle, just as they were earlier shown to depend on the public's acceptance of the causal and capacity principles.

The utilitarian view assimilates all punishment to mere penalties, designed only to secure compliance with the law. The lawbreaker is viewed no differently from the basketball player in "foul trouble" near the end of a game, who takes a calculated risk and "fouls out." The criminal has made a game-theoretic calculation based on the expected utility of compliance versus breaking the law. In this light, his decision to break the law is an error of calculation, rather than a moral fault.[2] To view murderers, robbers, arsonists, terrorists, etc., in this way contradicts our intuitions that there are clear

---

[1]Bentham allows that, aside from social defense, a "collateral end" of punishment is the "vindictive satisfaction" it gives to the victims of crimes, or to their friends and relatives. In this way, Bentham does provide a minor role for the expressive principle. However, this principle can only be accomodated as a by-product of utilitarian goals, since "the immediate principal end of punishment is to control action." An Introduction to the Principles of Morals and Legislation, Chapter XIII, section 1.

[2]This is the attitude toward criminals implied by Dennett's description of the deterrent effect of punishment. He argues that people "ought not to object to paying the assigned penalty" for "risk taking" and "gambles" which lose. See Elbow Room, pp. 164-5.

defense.] Yet its effectiveness as social control

differences between such criminals and, for example, sportsmen who commit fouls, people who are penalized for registering late for college courses, or parking violators. By not regarding the expression of condemnation as intrinsically valuable, utilitarianism violates the intuitions which underlie the distinction just described.

Viewing punishment as being partly justified as an expression of society's disapproval of the criminal also helps to obviate criticisms frequently made of retributivism. The proportionality principle, it will be recalled, is the innocuous-looking doctrine that punishment should be graded to correspond to the severity of the crime. However, this modest demand for proportionality between crime and punishment was taken to harsh extremes by Kant, who required "the principle of equality," i.e., "an eye for an eye," etc.[1] Even if we reject such a crude interpretation of proportionality (which would clearly be absurd, e.g., in the case of rape), problems remain in establishing a scale of measurement of the gravity of crimes and punishments. How can we decide, for example, whether 10 years in prison is a "fitting" punishment which "corresponds" to the crime of second degree murder? H.L.A. Hart points out a more general difficulty with assessments of the gravity of crimes. Does moral gravity depend on the

---

[1]Quoted by Edmund L. Pincoffs, The Rationale of Legal Punishment (Humanities Press, 1966), p. 2.

extent of harm caused, or is it rather a matter of subjective culpability (i.e., what degree of foresight, intention or malice was involved)? If it is the former, we have the problem of deciding how, without being barbaric, we can exact from the criminal the same harm which he inflicted on the victim. If it is the latter, how can any general scale of punishments take account of the widely differing circumstances which may be relevant to subjective culpability in each case?[1] (It must be said that the final objection is not conclusive. A category of crime is assigned a range of punishments, precisely so that a judge may allow for the particular circumstances of the case in passing sentence.)

The retributive principle of proportionality can be protected from these criticisms by shifting the emphasis away from the actual amount of hard treatment imposed, and stressing instead its symbolic significance. Proportionality can be achieved without reference to the rather crude "scales" metaphor which requires causing the criminal as much suffering as he inflicted on his victim. Instead, proportionality consists in the correspondence between our perception of the severity of the crime, and the amount of disapproval expressed by our punishment. It is conceivable that we should have an alternative method of

---

[1] See H.L.A. Hart, "Punishment and the Elimination of Responsibility," op. cit., pp. 161-3.

condemning criminals which would involve public denunciation rather than hard treatment.[1] However, as things stand, the infliction of hardship on criminals is the most recognized and convenient way of expressing our disapproval. A further justification for our use of hard treatment as our means of expressing disapproval is Feinberg's frank admission that that "vindictive resentment," and not just calm condemnation, is one of the motives behind punishment.[2] Visibly hard treatment is, more than mere public denunciation, likely to satisfy vengeful attitudes. This latter suggestion of Feinberg could be re-cast in a more palatable form in terms of what I called "the justice principle." The hard treatment of criminals can be justified as the restoration of fairness, in view of the law-abiding citizens who have made sacrifices to refrain from crime. The unfair advantage gained by criminals is more readily neutralized by hard treatment than by the mere expression of disapproval.

Nonetheless, although this "denunciatory theory of punishment" provides a defensible version of the proportionality principle, the theory has been criticized by Hart on the three grounds.[3] (Feinberg's defense of

---

[1] Feinberg, op. cit., pp. 115-6.

[2] Ibid., pp. 100-1.

[3] Op. cit., pp. 170-3.

condemning criminals which would involve public

the denunciatory theory came after Hart's attack, which was directed at its earlier exponents.) (1) Letting judges decide how to express society's disapproval of the criminal allows them too much discretion, and creates the risk that important social policy goals will be neglected in sentencing. (2) Our society is "morally a plural society," and judges may substitute their own moral judgments in the absence of a clear moral consensus in society as a whole. (3) The role of denunciation should be instrumental to social goals, and not an end in itself: "We do not live in society in order to condemn, though we may condemn in order to live." To reiterate what I stated at the beginning of this section, to give a comprehensive reply to objections such as these is beyond the scope of this brief defense of a retributive theory. Brief comments are in order.

The first two objections assume the undesirability of the influence of moral judgments on legal decisions, which should be based as far as possible on a strict application of existing rules. This central tenet of Hart's legal positivism is the center of much controversy in contemporary philosophy of law, and is not universally accepted. In particular, it has been comprehensively criticized by Ronald Dworkin, who argues that, especially in hard cases, moral and political judgments are part of

the law.[1] As for the third objection concerning the role of denunciation, there is the suspicion of a "straw man" argument. No one would seriously claim that this is the goal of society. What is claimed by retributivists is that denunciation is a worthwhile goal, which is intrinsically valuable, and Hart's argument does not refute this. All three criticisms exhibit the fear that regarding denunciation as an end in itself will result in the neglect of utilitarian goals. But it is surely possible to be sensitive to, and to further, deterrence, reform, etc., while at the same time regarding the expression of condemnation as a major purpose of punishment.

Two further criticisms of a purely utilitarian approach, as expressed by Dennett, have emerged from my own defense of compatibilism. First, I criticized Dennett's pragmatic approach, not to punishment, but to the hard determinist argument against moral responsibility. (See Chapter III, section 5.) Dennett argued that, so far as we finite beings will ever know, several alternative actions are always open to us at any given moment; and that this "epistemic openness" is sufficient justification for holding people morally responsible. Furthermore, holding ourselves and others responsible will have the beneficial

---

[1]Ronald Dworkin, Taking Rights Seriously (Harvard 1977). Chapters 1-5 are devoted to different aspects of this theme, but especially Chapter 2, "The Model of Rules I."

consequences of encouraging conformity to the law, and self-improvement. This can be viewed as a utilitarian justification of the institutions of moral praise and blame. My criticism of this approach was that it ignores doubts about the fairness of holding people responsible. Regardless of epistemic openness, and regardless of the benefits of holding people responsible, is it just to do so if all actions, including our attempts to form qualitative evaluations, have sufficient causal explanations? These doubts arise about the justice of our practices "in a sense not expressible in a purely utilitarian vocabulary" (above, p. 186). It is a weakness of Dennett's utilitarian approach that it cannot even describe these substantial doubts.

Second, I turned to Dennett's treatment of the problem of moral luck, which is the ultimate source of the hard determinist doubts about fairness described in the previous paragraph. Although his solution was, I argued, generally successful, his adoption of a utilitarian justification of punishment greatly weakens his argument. The Achilles heel of Dennett's treatment of moral luck is, I argued (Chapter IV, section 3), his admission that there is no principled justification for excusing some victims of bad moral luck, while holding others responsible. A detailed examination of the circumstances leading up to every crime would lead to universal exculpation, so we must arbitrarily refuse to

consequences of encouraging continuity to the [illegible]
self-improvement. [illegible]
justification [illegible]
blame. [illegible]
doors [illegible]

consider exculpating details beyond a certain point. To justify our practice of excusing certain categories of wrongdoers, Dennett relies upon the disutility of blaming children, insane people, etc. Crucially, these people no more deserve to be excused for their misdeeds than do rational, competent adults. This utilitarian approach to punishment and excuses violates our intuitions that the exculpation of some groups of offenders is required by fairness and justice. My own criterion of responsibility (Chapter III, section 3), on the other hand, respects and explicates these intuitions.

3. Why Most Compatibilists are Utilitarians

Hard determinists believe that the causation of human actions renders inapplicable and unjustified any attributions of moral responsibility.

> To say this is, of course, not to say that we should not punish criminals. Clearly, for our own protection, we must remove them from our midst so that they can no longer molest and endanger organized society. And, of course, if we use the word "responsible" in such a way that justly to hold someone responsible for a deed is identical with being justified in punishing him, then we can and do hold people responsible. But this is like the sense of "free" in which free acts are voluntary ones. It does not go deep enough. In a deeper sense we cannot hold the person responsible.[1]

---

[1] John Hospers, "Free Will and Psychoanalysis," Philosophy and Phenomenological Research, 1950, reprinted in Edwards and Pap (eds.), A Modern Introduction to Philosophy, 3rd edition (The Free Press, 1973), p. 89.

consider exonjbatiuñ Jeatiy pëJouq a cetetyn Jojuq. To
justify our bracetac ox cuwaiuñ certain ecceptiom ou
misunderstand, Jeatiy ... [illegible]
children, ... [illegible]
mole ecexiy ...

The rejection of moral responsibility leaves room only for a utilitarian justification of punishment, with social defense as its only goal. Retribution, with its references to giving the criminal what he deserves, and to our condemnation of him, cannot survive the elimination of moral responsibility.

Hard determinists share with libertarians the belief that "contra-causal freedom" is necessary for moral responsibility. Hard determinists reject this freedom, and hence reject responsibility. Libertarians, though, believe that we do indeed have such freedom, and are therefore not barred from holding a retributive theory of punishment. Compatibilists, on the other hand, agree with hard determinists by allowing that all actions may well be caused. However, the distinctive feature of compatibilism is the belief that causation does not rule out moral responsibility. In view of this feature, it is surprising that most compatibilists, despite their rejection of the need for contra-causal freedom, follow hard determinists in adopting a utilitarian theory of punishment. Compatibilists' belief in moral responsibility would seem to leave open the possibility of a retributive view, with its attendant moral notions (disapproval, desert, fairness, etc.).

So why have most compatibilists advocated a utilitarian theory of punishment? I suggest that it is because they

cannot find a satisfactory reply to the slippery slope argument, which I presented at the beginning of Chapter I as a graphic illustration of the determinism/moral responsibility problem. As we saw, the hard determinist argument begins by pointing out uncontroversial cases in which causation excludes moral responsibility. Consistent application of these excuses leads, if determinism is true, to exculpation for all misdeeds. To respond to this argument, compatibilists have to show how blameworthy and excusable actions can be distinguished, even if they are all caused. Their answer, as we saw, was that the distinction relevant to moral responsibility is not between freedom and causation, but rather between freedom and compulsion (Chapter I, sections 3a and b). We are responsible for those of our actions that are the natural, uncoerced result of our desires (even though such actions are, of course, caused).

However, modern compatibilists have realized that this distinction is not clear cut. Schlick, for example, struggles with the borderline cases of people under the influence of alcohol or narcotics.[1] Even though such people are freely acting on their desires, these desires themselves may exert compulsion. Encouraged by such cases, and further armed by the discovery of unconscious

---

[1]Moritz Schlick, "When Is a Man Responsible?" Problems of Ethics, reprinted in Edwards and Pap, op. cit., pp. 62-3.

cannot find a satisfactory reply to the 'dipper' stage
argument, which I discussed at the beginning of Section
as a graphic illustration of the criterion
responsibility
argument
which

motivation by modern psychologists, hard determinists such as Hospers return to the slippery slope argument. In the final analysis, he argues, all of our desires would be discovered, on close enough inspection, to compel us, hence rendering the compatibilist criterion useless.

In view of such difficulties in distinguishing compulsive from free, responsible behavior, it is not surprising that compatibilists sought a firmer basis for this distinction. I suggest that it was in search of a sharper distinction that they linked responsibility to a utilitarian theory of punishment. No longer do compatibilists have to produce elaborate and questionable distinctions between compulsive and free, natural desires. Instead, the responsible person is one on whom punishment will be effective by influencing his future motives in a desirable way. Thus responsibility becomes more of an empirical question, to be settled by gauging the likely effect of punishment on different offenders.[1]

My diagnosis applies particularly well to Dennett's reasons for adopting a utilitarian justification of punishment. (See my criticism of Dennett in Chapter IV, section 3, and my reference to this at the end of the last section.) Dennett effectively concedes to the hard determinist the main point in contention: the rationale

---

[1]Such accounts are given by Schlick, op. cit.; and Nowell Smith, "Free Will and Moral Responsibility," Mind, 1948.

for excusing people who are insane, under duress, deceived, too young, etc., can be extended to cover everyone who acts badly. A careful investigation into "the particular micro-details of the accused's circumstances" would lead us to excuse every criminal. There is no principle to justify blaming any particular criminal, while continuing to excuse those who traditionally qualify for exculpation. Dennett's response to this "subversive proposal," which plays into the hands of the hard determinist and his slippery slope argument, is to confine responsibility to purely utilitarian considerations. We need to maintain the "bracing effect" of our system of deterrence by drawing an "arbitrary . . . threshold for legal competence." Where we draw this line will be aimed at maximizing deterrence. It is Dennett's explicit admission that the slippery slope argument cannot be halted which forces him to turn to a utilitarian justification of our practices of blaming and excusing.

Still, this utilitarian justification does accord rather well with our current practices of allowing ignorance, compulsion, insanity, and extreme youth as excuses. Such people would very likely not respond to punishment. However, as has frequently been pointed out, the punishment of such people may be justifiable on the purely utilitarian ground that such a stern system would

deter other people.[1] Dennett, following other utilitarians, responds to this attempted reductio ad absurdum by pointing out the decrease in general respect for the law which is likely to result from such an unfair system.[2] Nonetheless, Dennett's reply is vulnerable to a criticism made in my general critique of utilitarian theories of punishment. He condemns unfairness only because of its adverse consequences in terms of falling respect for the law, whereas our intuitions regard it as intrinsically undesirable. As I have frequently pointed out, the criterion of responsibility I defend in Chapter III, in contrast, justifies our practice of excusing certain people, by showing that they deserve to be excused. In this way I explain why it is unfair to punish, for example, those not capable of obeying the law.

4. Why Compatibilists Need Not be Utilitarians

The preceding section illustrates rather well a criticism levelled at compatibilists by incompatibilists (both hard determinists and libertarians). The notion of moral responsibility which is permitted by the utilitarian view of punishment adopted by most compatibilists is a severely

---

[1] e.g., H.L.A. Hart, op. cit., pp. 17-21.

[2] Elbow Room, pp. 162-3.

restricted one. It is a pragmatic, makeshift substitute for responsibility, argue incompatibilists, which does not convey what we really mean when we hold people responsible. (Of course, the hard determinist believes that this makeshift substitute is the only feasible conception of responsibility.) This is another example of the general incompatibilist argument that compatibilism does not capture the full depth of our moral concepts, and instead only provides superficial substitutes. (See Chapter II, section B6.)

This "one-eyed utilitarianism" is the very weakness in most compatibilist accounts which Peter Strawson was concerned to correct in his famous paper "Freedom and Resentment."[1] It is interesting that Dennett, who refers approvingly to several of the points made by Strawson in this essay, proceeds to advocate the very utilitarian theory which is one of its major targets for criticism. Incompatibilists - whom Strawson calls "pessimists"[2] - are correct, admits Strawson, in claiming that there is a "lacuna" in the "optimists"'s (compatibilists') account of moral responsibility. The latters' utilitarian justification of our attributions of responsibility,

---

[1] Proceedings of the British Academy, 1962, reprinted in Watson (ed.), op. cit.

[2] Strawson also refers to "moral sceptics," who believe that moral concepts are empty whether or not determinism is true.

for [illegible]

in terms of the effectiveness of legal and moral sanctions, leaves out the heart of our moral framework: our "moral reactive attitudes."

> . . . these practices [i.e., praise/blame, and punishment/reward], and their reception, the reaction to them, really <u>are</u> expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes. . . . Indeed the very understanding of the kind of efficacy these expressions of our attitudes have turns on our remembering this.[1]

Much of "Freedom and Resentment" is devoted to a detailed analysis of the various kinds of "participant" or "reactive" attitudes, the most important of which, for our purposes, are <u>moral</u> attitudes. The point of this analysis is to show that it is these moral attitudes which are needed to fill in the lacuna in utilitarian, compatibilist accounts; but that the absence of determinism, which incompatibilists are wont to demand, <u>is neither necessary nor even relevant to our attributions of responsibility</u>. Having supplied the incompatibilist with what is missing in most compatibilist accounts, Strawson requires of the incompatibilist "a surrender of his metaphysics"[2] (i.e., his demand for contra-causal freedom).

The incompatibilist might well agree that the existence of moral reactive attitudes is the missing link which explains and justifies our institutions of blaming and

---

[1] Strawson, in Watson (ed.), op. cit., p. 80.

[2] Ibid., p. 78.

[illegible] of the effectiveness of Judge Hand's Learned Hand Application,

leaves out the heart of our [illegible] Court "Hand-

rective advice."

punishment. His point is rather that these attitudes themselves would be rendered unjustified by the truth of determinism. Determinism, it is feared, would leave room only for "objective attitudes," i.e., the kind of non-blaming, therapeutic attitude adopted towards criminals under a purely utilitarian system. In reply, Strawson admits that there are many occasions when we are indeed driven to abandon moral reactive attitudes and to adopt an objective standpoint. However, our doing so has nothing to do with a belief or fear that determinism is true. We abandon moral attitudes on precisely those occasions when traditional excuses operate: involuntary movements, ignorance, duress, infancy, insanity, etc. Most of the time, moral attitudes are perfectly possible and appropriate.

A final thrust by the incompatibilist remains: if we took seriously the implications of determinism, would we not have to extend the traditional excuses, and abandon moral attitudes in all cases? Strawson's reply is at once forceful and yet slightly unsatisfying. Reactive attitudes in general are an ineliminable part of our conceptual scheme, and there is no point in asking whether we should do what it would be impossible for us to do (i.e., abandon these attitudes).[1] These attitudes are a "given,"

---

[1] Ibid., pp. 74-5.

punishment. His point in raising that those activities
themselves would be ignored unnecessarily if the view of
deterrence were correct. [illegible] would [illegible]
only for "[illegible]" [illegible]
non-[illegible] [illegible]
[illegible]

without which human society would be unintelligible. Similar to induction, they neither can nor need to be justified.[1]

Strawson's point that the moral attitudes which are essential to our attributions of responsibility are not driven out by determinism can be made in a less dramatic and controversial way. He himself refers to cases where either objective or reactive attitudes are possible. For example, we may adopt an objective attitude to a perfectly normal adult, for the sole purpose of obtaining "a relief from the strains of involvement" which arise in reactive attitudes.[2] The idea that on some occasions one may have a choice of which attitude to adopt is reminiscent of Dennett's notion of non-competitive stances, or levels of description of human action, the choice between which is made on pragmatic grounds. The same arguments which Dennett gives for the compatibility of intentional and physical descriptions can be used to resist the suggestion that objective attitudes should drive out reactive attitudes. (See Chapter II, section A.) On other occasions, only one attitude is appropriate: we may only view insane people objectively, or as Dennett would say, we cannot reasonably adopt the intentional stance towards them. Just as Strawson argues that the universal abandonment of reactive attitudes would be unthinkable for

---

[1] Ibid., pp. 78-9. [2] Ibid., p. 69.

people who live in a society, Dennett points out that descriptions of all human actions in non-intentional terms would be practically impossible. Given that our choice of attitude is already made partly on pragmatic grounds, Strawson can justifiably appeal to the same pragmatic criteria to resist the argument that determinism should lead us to completely abandon moral attitudes.

The kernel of truth contained in the pessimists' (incompatibilists') account, I suggest, is that, as a matter of psychological reality, it is difficult to simultaneously adopt both a reactive and an objective attitude toward the same person. A psychotherapist will probably be less efficient in diagnosing and helping to resolve her client's conflicts if she becomes friendly with him, and allows reactive attitudes to interfere with her objective standpoint. The mistake made by pessimists is to conclude from this that objective attitudes take priority, and render reactive attitudes irrational. In most cases either attitude is possible, though it may be difficult to adopt them simultaneously. Neither attitude has primacy over the other: they may be psychologically incompatible, but logically they are perfectly consistent. Indeed, it would be irrational to be persuaded, because of the truth of determinism, to abandon reactive attitudes. According to the pragmatic criteria which are our guide in our choice of attitude, it is entirely appropriate that we adopt reactive attitudes in our human interactions.

Strawson's argument that the moral reactive attitudes which are essential to holding people responsible are not ruled out by determinism leaves the field open for compatibilists to hold a retributive view of punishment. Strawson himself is agnostic with respect to determinism (he claims that he doesn't understand it[1]), but his major concern is to show its irrelevance to the reactive attitudes on which moral responsibility in part depends. In fact, the reactive attitudes to which Strawson accords such an important place are particularly well suited to a feature of retributivism which I stressed in my defense of it: its function as an expression of society's condemnation of the criminal. Following Feinberg, I criticized utilitarians for being unable to give an account of condemnation as an intrinsically valuable goal of punishment. Strawson's arguments show that one can endorse this "expressive principle" of retributivism while still believing in determinism. More generally, to the extent that the other retributive goals expressed in the principles listed at the beginning of this chapter - giving the criminal what he deserves, being fair to law-abiding citizens, proportionality between crime and punishment, etc., - are viewed as dependent on, or consisting in, moral attitudes, Strawson's arguments further illustrate the consistency of retributivism with compatibilism.

---

[1]Ibid., p. 59.

which are elementary to learning. Some refinements are now
ruled out by determinists. After the first Chapter
comparing it to more recent writers, it
Strawson [illegible] "Freedom and [illegible]
(in Gardiner and [illegible] [illegible]
Sometimes [illegible]
[illegible]
[illegible]
[

I view Chapters III-V of this dissertation as, in part, a detailed working out of Strawson's visionary suggestions for improving compatibilism. Whereas compatibilists like Dennett have persisted in justifying our moral framework by reference to its social utility, I have argued in terms consistent with a retributive view. In Chapter III, I argued that the standard excuses from responsibility can be justified, not just because punishment of those who qualify for them would be inefficacious, but rather because it would be simply unfair. I supported the charge of unfairness by arguing for and applying a criterion to determine when people may justly be held responsible. In Chapter IV, I tried to solve the paradox which ensues when moral luck influences our assessments of people's actions. A utilitarian would point out the destructive consequences of allowing people to fall back on bad moral luck as an excuse for virtually any misdeed. Dennett insists that the only way to avoid this consequence is to set a pragmatically-determined threshold below which we will ignore the contribution of moral luck to people's actions. I, on the other hand, explained to what extent it is fair that various kinds of moral luck should influence our moral assessment. Again using my criterion of responsibility, I argued that on some occasions people deserve to be given extra "moral leeway" because of excessively bad moral luck. Finally, in the present chapter I have argued

directly in favor of a retributive theory of punishment, and pointed out that this is quite consistent with my compatibilist position.

A utilitarian approach to our attributions of responsibility, to the problem of moral luck, and to our practice of punishment, leaves room for only objective attitudes towards people. What I consider to be the most objectionable feature of this objective, utilitarian approach is that it fosters no difference in our attitude toward blameworthy and excusable people. Guilty people are punished, but only because of the beneficial consequences this will bring. We do not punish them because they deserve it, and neither do we excuse people because they deserve to be excused.

My criterion of responsibility, which also plays an important role in my treatment of moral luck and punishment, attempts to respect the intuitions which are violated by this utilitarian approach. The reason why ignorance, compulsion, and insanity are legitimate excuses is because they impair a person's capacity to act on justifiable qualitative evaluations. By the same token, the wrongdoer who does not qualify to be excused could and should have acted on the right qualitative evaluations. Both blameworthy and excusable people are getting what they deserve, in view of their capacities, according to a criterion which clearly distinguishes them. In this light,

we are perfectly justified in adopting moral reactive attitudes towards them. In particular, we are justified in condemning the blameworthy person. These are the attitudes which underlie many of the retributive principles listed in section 1. Whereas most compatibilists leave no room for these vital moral attitudes, my criterion of responsibility, in terms of the capacity for qualitative evaluations, allows for them. Hence, unlike that of most compatibilists, my position is consistent with a retributive view of punishment.

I have also given a diagnosis of why compatibilists have been drawn towards a utilitarian view of punishment. My suggestion was that it was because they have been unable to find a fair and clear-cut dividing line between blameworthy and excusable misdeeds. (This is precisely what is claimed by advocates of the slippery slope argument, which I have presented throughout as the greatest challenge to be faced by compatibilists.) Utilitarianism provides an easily determinable dividing line based on the effectiveness of punishment as a deterrent. The innovation of my dissertation in this regard was the presentation of a criterion of responsibility, based on the capacity to act on qualitative evaluations. This criterion provides a framework in which to determine whether, given the circumstances in any particular case, it is fair to hold the agent responsible. Because of this criterion, I do not

need to resort to utilitarian considerations to distinguish between blameworthy and excusable actions, and remain free to hold a retributive view.

In this regard (i.e., response to the slippery slope argument), I view my compatibilist arguments as an improvement over those of Strawson. He allows that, in the cases where we traditionally excuse people, an objective attitude tends to replace our moral attitudes. However, it is "practically inconceivable"[1] that we could be driven, by a belief in determinism or by anything else, to adopt an exclusively objective attitude to everyone. It is here that utilitarians, with their purely objective attitude, fail to do justice to our moral experience. Nonetheless, aside from his assurances that it is practically inconceivable that we could be persuaded to stop blaming people altogether, Strawson gives no criterion to help us determine exactly when we should excuse people, and adopt an objective attitude. The slippery slope from occasional exculpation to universal exculpation has not been convincingly answered.

A.J. Ayer has argued, in his discussion of "Freedom and Resentment," that Strawson never satisfactorily refutes the possibility of adopting a thoroughgoing objective attitude

---

[1]Ibid., p. 68.

to all human actions.[1] Given this possibility, there is a sneaking suspicion that, if determinism is true, it might be irrational to refuse to adopt this objective, scientific attitude. Strawson's insistence that moral reactive attitudes are ineliminable may be an ill-founded, sentimental attachment to a piece of folklore. Strawson may have shown that it is rational, in the sense of "prudent," to retain moral attitudes the abandonment of which would involve a chaotic upheaval of our conceptual scheme. But, as Ayer points out,

> there is another sense of "rational" in which the rationality of an attitude is measured not by the consequences of adopting it but by the standing of the beliefs which enter into it.[2]

In this sense, it may indeed be rational, if determinism is true, to abandon moral attitudes in favor of an objective outlook. It is curious to note that, in this light, Strawson may be offering pragmatic, utilitarian justifications for our use of retributive notions. Is this not the very instrumental view of morality against which Strawson is arguing in his criticisms of most compatibilists?[3] It is important to see that my proposal

---

[1] "Free Will and Rationality," Zak Van Straaten (ed.), Philosophical Subjects (Oxford, 1980), esp. pp. 9-13.

[2] Ibid., p. 11.

[3] See the long quotation from "Freedom and Resentment," above.

of a criterion of responsibility halts the slippery slope, and protects me from this objection.

Despite this criticism, the colossal value of "Freedom and Resentment" remains its pioneering role in liberating compatibilism from a purely utilitarian justification of punishment and blame. My major goal in this chapter has been to follow the direction suggested by Strawson, and to argue that compatibilism leaves room for retribution. I first outlined utilitarian and retributive theories of punishment (section 1), and then offered a critique of the utilitarian view (section 2). I suggested that the reason why compatibilists have tended to be utilitarians is that they have failed to come up with an adequate reply to the slippery slope argument (section 3). Finally, in the present section (section 4) I have shown why compatibilists can be retributivists: (1) Strawson's arguments prove the general consistency of these two doctrines with each other. (2) My own version of compatibilism gives a satisfactory answer to the slippery slope argument, and hence obviates the need for recourse to utilitarian arguments.

and **criterion** **of** **theme** **within** **a** **and** **Recognition**," **criterion** **of** **theme** **within** **a**

and **pictures** **me** **from** **him** **objection.**

Despite this **criterion** **of** **theme** **within** **a** **and** **Recognition**," **criterion** **of** **theme** **within** **a** **and** **pictures** **me** **from** **him** **objection.**

BIBLIOGRAPHY

BIBLIOGRAPHY

Andre, Judith. "Nagel, Williams, and Moral Luck." Analysis, 1983.

Aristotle. Nichomachean Ethics.

Austin, J.L. "Ifs and Cans," Philosophical Papers, Oxford University Press, 1961.

_____. Sense and Sensibilia. Oxford University Press, 1962.

Ayer, A.J. "Freedom and Necessity." Philosophical Essays, Macmillan, 1954.

_____. "Free Will and Rationality." In van Straaten, 1980.

Ayers, M. The Refutation of Determinism: An Essay in Philosophical Logic. Methuen, 1968.

Benn, Stanley I. "Punishment." In Edwards, Paul (ed.), The Encyclopedia of Philosophy, Macmillan, 1967.

Bennett, Jonathan. "Accountability." In van Straaten, 1980.

Berofsky, Bernard (ed.). Free Will and Determinism. Harper and Row, 1966.

Bridgman, Percy W. "Determinism in Modern Science." In Hook, 1961.

Campbell, C.A. "Is 'Free Will' a Pseudo-Problem?" Mind, 1951, reprinted in Edwards and Pap, 1973.

Coyne, Margaret Urban. "Moral Luck?" Paper presented at Eastern Division American Philosophical Association meeting, Boston, December 1983.

Dennett, Daniel C. Brainstorms: Philosophical Essays on Mind and Psychology. Bradford Books, MIT, 1978.

_____. "Conditions of Personhood." In Brainstorms.

_____. "Intentional Systems." In Brainstorms.

_____. "Mechanism and Responsibility." In Brainstorms.

_____. "On Giving Libertarians What They Say They Want." In Brainstorms.

_____. "The Abilities of Men and Machines." In Brainstorms.

_____. Elbow Room: The Varieties of Free Will Worth Wanting. Bradford Books, MIT, 1984.

Dworkin, Ronald. Taking Rights Seriously. Harvard University Press, 1977.

Edwards, Paul. "Hard and Soft Determinism." In Hook, 1961.

Edwards, Paul and Arthur Pap (eds.) A Modern Introduction to Philosophy. 3rd edition, The Free Press, 1973.

Feinberg, Joel. Doing and Deserving. Princeton University Press, 1970.

_____. "Causing Voluntary Actions." In Doing and Deserving.

_____. "On Being 'Morally Speaking a Murderer.'" In Doing and Deserving.

_____. "Problematic Responsibility in Law and Morals." In Doing and Deserving.

_____. "Sua Culpa." In Doing and Deserving.

_____. "The Expressive Function of Punishment." In Doing and Deserving.

_____. "What Is So Special About Mental Illness?" In Doing and Deserving.

Feinberg, Joel and Hyman Gross. Philosophy of Law. 2nd edition, Wadsworth, 1980.

Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility." Journal of Philosophy, 1969.

_____. "Coercion and Moral Responsibility." In Honderich, T., ed., Essays on Freedom of Action, Routledge and Kegan Paul, 1973.

_____. "Freedom of the Will and the Concept of a Person." Journal of Philosophy, 1971, reprinted in Watson, 1982.

Gallois, Andre. "Van Inwagen on Free Will and Determinism." Philosophical Studies, 1977.

Gaylin, Willard. The Killing of Bonnie Garland. Penguin, 1983.

Ginet, Carl. "Can the Will Be Caused?" Philosophical Review, 1962.

Goldman, Alvin. "Actions, Predictions, and Books of Life." American Philosophical Quarterly, 1968.

Hare, R.M. "Nothing Matters." Applications of Moral Philosophy, Macmillan, 1972, reprinted in Klemke, 1981.

Hart, H.L.A. Punishment and Responsibility: Essays in the Philosophy of Law. Oxford University Press, 1968.

_____. "Changing Conceptions of Responsibility." In Punishment and Responsibility.

_____. "Legal Responsibility and Excuses." In Punishment and Responsibility.

_____. "Murder and the Principles of Punishment: England and the United States." In Punishment and Responsibility.

_____. "Prolegomenon to the Principles of Punishment." In Punishment and Responsibility.

_____. "Postscript: Responsibility and Retribution." In Punishment and Responsibility.

_____. "Punishment and the Elimination of Responsibility." In Punishment and Responsibility.

Hart, H.L.A. and A.M. Honore. Causation in the Law. Oxford University Press, 1959.

Hermann, Donald. The Insanity Defense: Philosophical, Historical, and Legal Perspectives. Thomas, 1983.

Hick, John. Philosophy of Religion. Prentice-Hall, 1963.

Hobart, R.E. "Free Will as Involving Determinism and Inconceivable Without It." Mind, 1934, reprinted in Berofsky, 1966.

Hook, Sidney. Determinism and Freedom in the Age of Modern Science. Collier Books, 1961.

Hospers, John. "Free Will and Psychoanalysis." Philosophy and Phenomenological Research, 1950, reprinted in Edwards and Pap, 1973.

Hume, David. An Enquiry Concerning Human Understanding. 1748.

James, William. "The Will to Believe." Essays in Pragmatism, Hafner, 1948.

Kant, Immanuel. The Metaphysical Elements of Justice. 1797.

Kenny, Anthony. Freewill and Responsibility. Routledge and Kegan Paul, 1978.

Klemke, E.D. The Meaning of Life. Oxford University Press, 1981.

Malcolm, Norman. "The Conceivability of Mechanism." Philosophical Review, 1968, reprinted in Watson, 1982.

Mill, John Stuart. A System of Logic. 1843. The relevant section is reprinted in Edwards and Pap, 1973.

_____. An Examination of Sir William Hamilton's Philosophy. 1865. The relevant section is reprinted in Berofsky, 1966.

_____. On Liberty. 1859.

_____. Utilitarianism. 1863.

Morris, Herbert. "Persons and Punishment." The Monist, October 1968, reprinted in Feinberg and Gross, 1980.

Moore, Michael S. "Legal Conceptions of Mental Illness." In Brody, B.A. and H. Tristram Englehardt (eds.), Mental Illness: Law and Public Policy, Reidel, 1980.

Nagel, Thomas. "Moral Luck." Mortal Questions, Cambridge University Press, 1979.

Narveson, Jan. "Compatibilism Defended." Philosophical Studies, 1977.

Nowell Smith, P.H. "Free Will and Moral Responsibility." Mind, 1948

Pincoffs, Edmund L. The Rationale of Legal Punishment. Humanities Press, 1966.

Quine, Willard V.O. Word and Object. The MIT Press, 1960.

Rorty, Richard (ed.). The Linguistic Turn. Chicago, 1967.

Ryle, Gilbert. The Concept of Mind. Hutchinson, 1949.

Salmon, Wesley. "Determinism and Indeterminism in Modern Science." In Feinberg, Joel (ed.), Reason and Responsibility, 4th edition, Dickinson, 1978.

Schlick, Moritz. "When Is a Man Responsible?" Problems of Ethics, Prentice-Hall, 1939, reprinted in Edwards and Pap, 1973.

Singer, Peter. Practical Ethics. Cambridge University Press, 1979.

Smart, J.J.C. "Sensations and Brain Processes." Philosophical Review, 1959, reprinted in Edwards and Pap, 1973.

Smart, J.J.C. and Bernard Williams. Utilitarianism: For and Against. Cambridge University Press, 1973.

Strawson, Peter F. "Freedom and Resentment." Proceedings of the British Academy, vol. xlviii, 1962, reprinted in Watson, 1982.

Suter, Ronald. Are You Moral? University Press of America, 1984.

Taylor, Charles. "Responsibility for Self." In Rorty, Amelie O., ed., The Identities of Persons, University of California Press, 1976, reprinted in Watson, 1982.

Taylor, Richard. Good and Evil. Macmillan, 1970.

_____. "The Meaning of Human Existence." In Leiser, Burton, ed., Values in Conflict: Life, Liberty, and the Rule of Law, Macmillan, 1981.

Trusted, Jennifer. Free Will and Responsibility. Oxford University Press, 1984.

Van Inwagen, Peter. "Ability and Responsibility." Philosophical Review, 1978.

_____. An Essay on Free Will. Oxford University Press, 1983.

_____."Reply to Gallois." Philosophical Studies, 1977.

_____."Reply to Narveson. Philosophical Studies, 1977.

_____. "The Incompatibility of Free Will and Determinism." Philosophical Studies, 1975, reprinted in Watson, 1982.

van Straaten, Zak, ed. Philosophical Subjects: Essays Presented to P.F. Strawson. Oxford University Press, 1980.

Watson, Gary. "Free Agency." Journal of Philosophy, 1975, reprinted in Watson, 1982.

_____, ed. Free Will. Oxford University Press, 1982.

Williams, Bernard. Problems of the Self. Cambridge University Press, 1973.

_____. "Ethical Consistency." In Problems of the Self.

_____. "Morality and the Emotions." In Problems of the Self.

_____. Moral Luck. Cambridge University Press, 1981.

_____. "Moral Luck." In Moral Luck.

_____. "Utilitarianism and Moral Self-Indulgence." In Moral Luck.

1983.

MICHIGAN STATE UNIVERSITY LIBRARIES
3 1293 03071 4814