“OUGHT” CLAIMS AND BLAME IN A DETERMINISTIC WORLD

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

James Austin Hobbs

January 2013

“Ought” Claims and Blame in a Deterministic World

James Austin Hobbs, Ph. D.

Cornell University 2013

This dissertation examines two aspects of morality: (i) normative “ought” claims and (ii) blame. It is argued that these aspects of morality are compatible with “determinism,” the theory that all events are causally necessitated by prior events and the laws of nature. No position is taken on the likelihood that the world is deterministic. Instead, this dissertation rebuts arguments for incompatibilism and attempts to explain why determinism, if true, should not undermine these two aspects of morality.

Chapter One focuses on the intuitive, but controversial, proposition that moral “ought” claims imply “can” claims. This principle can seem to entail that, if determinism is true, no one ought to do anything other than what he or she actually does. Chapter One argues that “A ought to do X” implies only that A has the ability and the opportunity to do X, not the will to do X, and thus, that it can be true that a person ought to do something even if that person is causally determined to do otherwise.

Chapters Two and Three focus on blame and blameworthiness. Chapter Two reviews contemporary accounts of blame and offers an alternative account, on which blameworthiness involves a change in moral standing, including a loss of claims against certain forms of suffering and the acquisition of special obligations. Blaming involves actually caring about these

changes by treating a person with decreased moral respect and demanding that the special obligations are satisfied.

Chapter Three addresses the idea that a person cannot be blameworthy for an action unless she could have done otherwise. Many who believe that moral responsibility is compatible with determinism reject this principle on the basis of a counterexample developed by Harry Frankfurt. Chapter Three argues for a conception of moral responsibility that entails a more plausible principle about responsibility: that to be blameworthy, it must be possible for the person to have behaved differently. This weaker principle is not shown to be false by Frankfurt's counterexample, it is necessary to support judgments about the desert of blame, and it does not lead to incompatibilism about determinism and moral responsibility.

BIOGRAPHICAL SKETCH

James Hobbs was born in Kingston, New York, the son of Pamela Newcombe Hobbs (now Abele) and William Henry Hobbs IV. James began his education at the Washington Irving Elementary School in Catskill, New York. He completed secondary school at the Albany Academy in Albany, New York, and his undergraduate education at Reed College, in Portland, Oregon. James received a Master of Arts in Philosophy from Cornell University, and then attended New York University School of Law, where he earned his Juris Doctor degree. Following a clerkship with the Honorable Janet C. Hall, United States District Judge, and several years of legal practice, James returned to Cornell to submit his dissertation and earn his Doctorate in Philosophy.

For my brother Bill.

ACKNOWLEDGMENTS

I would like to thank Cornell University, the Sage School, and its faculty, and especially my present and former committee members for their support and encouragement, their wonderful insights, and their flexibility in allowing me to complete this project.

I owe the biggest debt to my family who always encouraged me, and especially to my wife and my father for insisting that I finish.

TABLE OF CONTENTS

INTRODUCTION

The natural sciences tend to suggest—in their assumptions and as an inference from their success—that all events can be explained in terms of the behavior of matter and energy in accordance with natural laws. Human actions are not exempt from such explanations. Indeed, advances in neuroscience, physiology, and related fields consistently expand our ability to explain human action in terms of physical and chemical events occurring in accordance with natural laws. These advances appear to support the idea that in principle all human actions could be explained as the effect of chemical and physical processes occurring principally within our brains.

The possibility of a thoroughgoing explanation for human behavior in terms of physical events and physical and chemical properties of our brains is in tension with the way we view ourselves when we contemplate acting. As we see it, what we do depends on what we think, what we care about, what we believe, and what we want. All of these are attitudes that, unlike neurological events, we can experience through self-consciousness and are nowhere to be found in a brain scan or neurological diagram. The thought that our behavior can be explained as the result of a complex series of brain processes — processes beholden to the laws of chemistry and physics and outside the scope of our self-consciousness — threatens our sense that our choices are dictated by thought, deliberation, and choice – processes of which we are, more or less, aware and which are subject, if at all, to norms such as

logic, reason, “common sense,” and morality, not to natural laws. Thus, we must ask whether our scientific understanding of the world can be squared with our conception of ourselves as agents.

This question can arise in different forms and deal with a number of different aspects of our conception of ourselves as agents. Perhaps the most fundamental and far reaching of the problems evoked by the naturalistic, physicalist view of the world is the mind-body problem. If there are physical explanations of our behavior, which make reference to neurons and brain chemistry, not beliefs, desires, and other mental attitudes, then we are left with the problem of explaining the very nature of these mental attitudes, how they relate to the physical world, and what role if any they play in causing or explaining our behavior. Though these are fascinating problems, they are not the subject of this dissertation. I must assume here that these problems can be resolved in a way that supports our natural tendency to speak of beliefs, desires and other mental attitudes as explanations of our actions. Specifically, I assume that regardless of whether or not it is true to say that I did X because of the firing of certain neurons moments beforehand, it is also meaningful, and sometimes true, to say that I did X because, for example, I wanted Y and I believed that doing X was the way to get Y or because I had been intending to do X and I believed this to be a good opportunity to carry out that intention.

Assuming that our behavior can often be explained in terms of mental attitudes or mental events, we are still left with another set of problems. For even if we may meaningfully and correctly speak of our actions as the product

of our beliefs, desires, values, principles, etc., it still may be that these attitudes and mental attributes themselves are determined by the physical and chemical events going on within our brains and bodies. This possibility can seem to undermine our sense of freedom, control, and choice over what we do. In short, the assumption that mental explanations are meaningful and potentially true does not allow us to ignore the physical explanations and the evidence that physical events have physical causes. These physical explanations can make the control that we appear to have over what we do seem to be an illusion. While it may seem to you that it is up to you whether to continue reading, stop for a break, or skip ahead, the scientific picture seems to suggest that what you do is settled by the behavior of physical and chemical particles within your brain, which has long been determined by natural laws. Thus, even assuming that there is a place for mental attitudes in the physical world, we are still left with the problem of explaining our sense that we are agents empowered with free wills.

The scientific picture of the world raises the classic problem of freedom and determinism. This problem might be expressed as an inconsistent triad: people sometimes act freely; the world is deterministic; and if the world is deterministic, then no one ever acts freely. Expressing the threat to freedom as “determinism” means that we are supposing that the laws of nature leave no room for indeterminism and that the state of the world at any given moment, taken together with the laws of nature, is sufficient to determine the

state of the world at any later moment.[1] This problem, like the mind-body problem, is also not the subject of this dissertation. However, unlike the mind-body problem, I need not assume a solution to this problem. Instead, I will set the problem aside and consider a third set of issues that arise from the possibility of a deterministic physical world.

The idea that we sometimes act freely is not the only aspect of our agency that is under threat from determinism. If what we do can be explained entirely in terms of deterministic laws of physics and chemistry, then it can be hard to see how we can salvage our conception of ourselves as *moral* agents, agents who are subject to moral norms and who deserve blame and punishment or praise and reward for our performance relative to those norms. Determinism raises a number of questions about moral agency, including the following: Can it be true that I ought to do X if the laws of physics and the history of the world entail that I cannot do X? And how, in that case, could I deserve blame for not doing X? These questions identify the two central problems of moral agency and determinism that are the subject of this dissertation.

The problems that determinism raises for moral agency are not always separated from the problems it raises for free will. It is common to assume

[1] Somewhat ironically, the natural sciences, which seem to raise the threat of determinism, may not in fact support a thoroughgoing determinism, at least not at the subatomic level. As I understand it, quantum mechanics holds that there is an element of indeterminacy in the behavior of subatomic particles. However, as I understand it, quantum indeterminacy does not carry over and create significant indeterminism in the behavior of atomic or "super"-atomic entities. Therefore, I assume that determinism remains a real possibility for entities above the subatomic level and that this possibility is supported by the natural sciences and the advances in neurology from which I began.

that the problems are bound together and that moral agency simply stands or falls with freedom. For example, Kant seems to have taken it as obvious that when we view the world as deterministic there simply is no room for moral claims of any sort.[2] Kant takes the possibility that we are moral agents to be entirely dependent on the possibility of metaphysical freedom, and thus, defending that possibility is a central theme running through the first two Critiques and the Groundwork. Kant is not alone. It remains common to collapse the problems that determinism presents for morality into the problem of free will. This is particularly true with respect to moral responsibility. It is often assumed that freedom is necessary (if not also sufficient) for moral responsibility, and those concerned about moral responsibility therefore often focus on the question of whether freedom is compatible with determinism.

In suggesting a departure from this approach, I do not claim that moral agency is entirely independent of freedom. There may be a connection between the two. Instead, I insist only that any dependence of moral agency upon free will is not obvious and not something that should be taken for granted. More importantly, I argue that we can make progress in evaluating the compatibility of moral agency and determinism while setting the problem of freedom and determinism to the side.

---

[2] For example, see the *Critique of Pure Reason* (Kant (1781/1787)), at A547/B575, where, using "nature" to mean a specifically deterministic conception of the world, Kant says: "indeed, the ought, if one has merely the course of nature before one's eyes, has no significance whatsoever."

One reason to treat the problems of moral agency as distinct and to approach them on their own terms is that it helps us avoid a philosophical minefield. It is an understatement to say that the ideas of freedom and free will are contested. Intractable may be the best way to describe the disagreement about what it means to act freely. From a tradition of philosophers and theologians running, at least, from Aristotle to Hume and Schopenhauer, we inherit a number of competing conceptions of freedom. This tradition distinguished a freedom that required only the absence of obstacles and thus the ability to do what we want (sometimes called the "liberty of spontaneity") from one which requires the ability to choose among truly open alternatives and thus presupposes indeterminism (sometimes called the "liberty of indifference"). Philosophers in this tradition have disagreed for millennia, about which, if either, of these senses of freedom should matter. As far as I am aware, few now accept a simple version of either sort of conception, but the 20th and 21st century debate can still be seen as a continuation of the ancient one. Harry Frankfurt's view that freedom of the will consists in having a will that is consistent with one's higher-order desires can be seen as a sophisticated, modern derivative of the position that it is the "liberty of spontaneity" that matters.[3] Robert Kane, who defines free will as "the power of agents to be the ultimate creators (or originators) and sustainers of their own ends or purposes" and suggests that this means that it must be "truly up to us what we do" and that we must be "the ultimate, buck-stopping

[3] *See* Frankfurt (1971).

originator[s] of our actions," can be seen as carrying forward the intuitions of those who insisted on the "liberty of indifference."[4] The persistence of this kind of disagreement about the very notion of freedom leaves me doubtful that we are likely to find enough common ground if we take the traditional, metaphysical approach to the problems of moral agency.

Nearly half a century ago, in a statement that remains apt, P.F. Strawson summarized the result of approaching moral responsibility as a metaphysical question about freedom of the will:

> [Incompatibilists say] that *just* punishment and *moral* condemnation imply moral guilt and guilt implies moral responsibility and moral responsibility implies freedom and freedom implies the falsity of determinism. And to this [compatibilists] are wont to reply in turn that it is true that these practices require freedom in a sense, and the existence of freedom in this sense is one of the facts as we know them. But what 'freedom' means here is nothing but the absence of certain conditions the presence of which would make moral condemnation or punishment inappropriate.
>
> * * *
>
> [The incompatibilist,] being the sort of man he is, has only one more thing to say; that is that the admissibility of these practices, as we understand them, demands another kind of freedom, the kind that in turn demands the falsity of the thesis of determinism.[5]

As Strawson illustrates, the move from moral responsibility to freedom tends only to generate a debate about which kind of freedom is required for moral

---

[4] Kane (1996) p.4.

[5] Strawson, P.F. (1962) pp.73-75.

responsibility.

Among those who focus on the compatibility of freedom and determinism, one occasionally encounters a tacit admission that the concepts of freedom and free will are unlikely to provide a toehold on issues of moral responsibility. Recognizing disagreement about the concept of free will, they turn back to the concept of moral responsibility to find purchase. For example, Al Mele explains:

> [Various uses of the term 'free'] do not concern me. My interest is in what might be termed *moral-responsibility-level free action*—roughly, free action of such a kind that if all the freedom-independent conditions for moral responsibility were satisfied without that sufficing for the agent's being morally responsible for it, the addition of the action's being free to this set of conditions would entail that he is morally responsible for it.[6]

But if this is how one is supposed to get a grip on the concept of free action, then why should we who are concerned with morality and moral responsibility bother with the problem of free will? It can seem as though we are told that if we want to know whether responsibility is possible in a deterministic world, we should ask whether freedom is possible, but if we point out that many different ideas are given the name "freedom," we are told to rely on the concept of responsibility to identify the concept of freedom at issue.[7] Thus, the traditional

---

[6] Mele (2006) p.17.

[7] My argument here is, admittedly, a bit quick. Many would claim that there are ways of getting a grip on the relevant notion of freedom without simply circling back to responsibility. On one important approach, we might understand both freedom and moral responsibility by analyzing and developing a notion of "control" over one's own actions. There may be some promise to this approach, but it cannot be claimed that the notion of control at issue here is

approach that assumes freedom is the key to moral agency appears to be a potentially unnecessary detour through treacherous terrain.

In addition, focusing on the problems of moral agency and determinism as independent problems allows us to keep these issues framed as, fundamentally, issues of moral theory, rather than metaphysics. Above, I phrased the question of the compatibility of determinism and moral responsibility as a normative issue: could anyone deserve to be held responsible if determinism is true? And the question of whether it makes sense to say that I ought to do something, even if determinism is true, can similarly be phrased as a question about the nature and scope of our moral obligations: to what extent do we have obligations to do things that we are unable to do? On an approach that assumes that freedom is the essential condition for moral agency, these moral questions are easily supplanted by a metaphysical issue: whether free will is possible in a deterministic world. The approach I take here remains focused on these moral questions and examines the moral phenomenon at issue closely in order to help address these issues. This is appropriate, as we should not conclude that morality cannot be reconciled with determinism without looking closely at our moral judgments and practices and carefully examining their purpose and conditions.

---

itself self-evident or safe from controversy. Thus, it is not obvious that this approach avoids unnecessary disputes. Moreover, unless we have a clear grip on what moral responsibility is and what its conditions are, we should not assume that "control" really is the relevant requirement for moral responsibility. And even if it is, we should not assume we are in a position to develop the proper understanding of control without first clarifying the idea of responsibility.

Of course this does not mean we can or should avoid all metaphysical questions. Ethics cannot be divorced from metaphysics entirely, particularly not in this area. But I seek to keep the focus on the moral phenomena at issue and to delve into metaphysical questions about ability, choice, or control only where necessary to understand these moral phenomena.

I cannot hope to address all of the questions that determinism raises for moral agency in this dissertation. Instead, I focus on two moral principles that raise central issues for the compatibility of determinism and moral agency. In Chapter One, I address the principle that "ought" implies "can," as this principle raises a clear problem about the compatibility of determinism and moral obligation. This principle limits the scope of what we ought to do to the range of actions that we are able to perform. Thus, when combined with determinism, it appears to drastically limit the scope of our obligations. If determinism means that there is only one course of action open to us at any time, then the "'ought'" implies "'can'" principle and determinism appear to entail that it is never the case that we ought to do anything other than what we in fact do. In other words, no matter how badly you act, it will never be true to say that you ought to have done otherwise. If this is right, then at any given time, either there is nothing you ought to do or what you ought to do is whatever you will in fact do. This conclusion would seem to undermine the entire point of moral beliefs by leaving us with little to say in terms of evaluating courses of action. I argue that this is a consequence we can avoid. By focusing on the significance and purpose of normative ought claims and on

our intuitions about particular cases, we can not only understand the appeal of the idea that “ought” implies “can,” but also formulate a more precise version of that principle. This version of the principle shows that determinism does not threaten the truth of common sense normative “ought” claims.

The second principle that I address is the principle that responsibility for an action, particularly praiseworthiness and blameworthiness, depends on having the ability to do otherwise. Given this principle, determinism appears to be incompatible with moral responsibility because determinism makes it impossible to do anything other than what we do. I do not turn to a discussion of this principle until Chapter Three. First, Chapter Two contains a discussion of the nature of blame and praise, including critical evaluation of various contemporary accounts of blame and argument in favor of my alternate proposal. The analysis of blame and praise in Chapter Two places us in a better position to understand what must be the case in order for a person to deserve blame or praise. Relying on this analysis, I argue that desert of blame and praise does not depend on the ability to do otherwise. It requires less than this. Praise or blame can be deserved just in case our actions reveal our moral quality, and this can be the case even where we cannot actually do anything other than what we do. Therefore, in Chapter Three, I argue for a principle that is weaker than the principle of alternate possibilities, and which does not support incompatibilism about determinism and moral responsibility.

These Chapters do not address every argument that might be taken to show that determinism is incompatible with moral obligations and with moral

responsibility. However, they defuse two of the most important arguments for this incompatibilism, and they provide at least some explanation for why these moral phenomena are compatible with determinism. At the end of Chapter Three, I close with some thoughts on the limits of this explanation and the possibilities for addressing its shortcomings.

CHAPTER ONE

WHAT CAN OUGHT IMPLY?

I. What a Person Ought to Do Depends on What She Can Do:

Years ago, during a winter walk in one of Ithaca's gorges, my dog Madison cut her paw rather severely. Typically, I would have taken Madison to my friend, Sue, who is a local veterinarian, but at the time, Sue was on vacation. She would not be back for a week. So, Madison and I went to the local veterinary hospital. The next time I saw Sue, I related the incident and complained about how much the hospital charged me. I expected her to say something like, "It's too bad I wasn't around." So, it was a bit of a surprise when she said, "You should have come to me. I wouldn't have charged you." As it turned out there was a miscommunication. I had mistakenly told Sue that the incident occurred in the prior week, rather than two weeks prior while she was gone. But suppose for a moment that there had not been a misunderstanding of this sort. It would have been hard to make much sense of her comment. If she had been aware of the facts, and yet she still thought that I ought to have had her take care of Madison, then she would seem to be committed to the idea that although it was impossible[8] for me to take my dog to her, I should have done so anyway. This is an idea that strikes me as

[8] I am employing a non-technical sense of "impossible" for now. Presumably there was nothing logically or metaphysically impossible about me taking Madison to see Sue. In fact, it may have been merely impractical or self-defeating to try. But let's suppose that Sue was at an exclusive resort on a distant island country; that all affordable means of travel were sold out; and that I would not have been permitted to bring my dog into the country or into the resort where Sue was staying.

obviously false, so obviously false that it must reflect confusion about what it means to say someone ought to do something. What sense is there in saying that a person ought to pursue some course of action that is not open to them?

I am inclined to see this example as an instance of a relationship gestured at by the phrase, "'ought' implies 'can'." Certainly, this example alone does not prove that "ought" implies "can." But it does suggest that *some* forms of impossibility, or *some* kinds of obstacles to performing an action, make it false to say that an agent ought to perform that action. Moreover, insofar as Sue's claim would involve confusion about what it means to say that someone should have done something, it suggests that there is a conceptual relationship between *some* "should" or "ought" claims and *some* claims about what is possible. It seems that, at least sometimes, what we ought to do depends on what we can do.

At the same time, there are significant problems for any simple understanding of the principle that "ought" implies "can." As will be discussed below, there are a variety of counterexamples to the broader formulations of the principle. In addition, when broadly stated, the principle seems to support an argument that would systematically undermine the truth of a huge class of common sense normative claims. I will discuss these problems in depth throughout this Chapter, and I will defend a version of the principle that avoids them while capturing common sense intuitions like those that drive my example here.

II. The Argument from Determinism:

There is a simple argument that shows the importance of developing a proper understanding of the relationship between "ought" and "can." According to this argument the truth of both determinism and the principle that "ought" implies "can" would make false a great number of highly plausible moral claims. For now, take determinism to be the thesis that, at any point in time, only one course of events from that time forward is consistent with the laws of nature conjoined with a complete description of the state of the world at that time. The argument can be put in the following way:

1) If determinism is true (and people cannot change the laws of nature, and people cannot change the past in any significant way), then no person can ever do anything other than what he or she actually does.

2) If a person ought to do something, then she can do that thing.

3) So, if determinism is true, then it is never the case that a person ought to do something other than what she, in fact, does.[9]

---

[9] Something roughly like this argument has appeared in print a number of times, and a number of philosophers have endorsed versions of the reasoning it represents. For example, both C.D. Broad and Ishtiyaque Haji seem to be committed to arguments of roughly this form. In "Determinism, Indeterminism, and Libertarianism" (Broad (1952)), Broad endorses the idea that "ought" implies "can," and uses it to argue that many moral "ought" claims are incompatible with determinism. Focusing on claims like "You did A, but you (morally) ought to have done not-A," Broad argues that for such claims to be true, it must be the case that the agent was capable of doing not-A in a sense that is incompatible with being physically determined to do A. So, though he never puts the argument in precisely the form I do, he seems to be committed to each of its premises, and its conclusion.

Haji (1999) argues for a stronger conclusion than the one I am considering. He believes that all moral ought claims, whether about actions one actually performs or not, are incompatible with determinism. He endorses the following principle: An agent has a moral obligation not to perform an action, only if she can refrain from performing it. Haji (1999) p.183. And he believes that if determinism is true and one does X then one cannot refrain

The argument appears to show that determinism is incompatible with a great number of claims that strike most of us as obviously true. According to this argument, if determinism is true then it would never be true to say, "You really shouldn't have lied to her like you did." It would never be true to say that a person who committed a heinous crime should have done otherwise. And of course, if this argument stands, then the truth of determinism would falsify the overwhelmingly plausible claim that Nazis should not have run concentration camps. This is an extraordinary and, at least to my mind, unacceptable conclusion. So, I shall be assessing the idea that "ought" implies "can" with this argument in mind and with the aim of showing that its conclusion can be

---

from doing X in the relevant sense. Haji (1999) p.188. So, he is committed to the conclusion that if determinism is true and an agent does X, then it is not true that she ought to have refrained from doing X.

Paul Saka and Derk Pereboom also seem committed to the validity, if not the soundness, of something quite like this argument. Saka (2000) offers a modal-epistemic version of the argument above, but he uses it as a *reductio* of the principle that "ought" implies "can." While his argument differs from the one above, it requires that determinism and "'ought' implies 'can'" entail the conclusion above. Pereboom (2001) discusses Haji's argument from determinism and "'ought' implies 'can.'" There, he accepts the idea that determinism and "'ought' implies 'can'" would undermine all sorts of central moral ought claims. He discusses, but does not fully recommend the option of denying "'ought' implies 'can.'" Pereboom (2001) pp.147-48. Instead, he seems to believe that he can regard these "ought" claims as false, while attempting to preserve the truth of judgments about goodness and badness, rightness and wrongness. *See* Pereboom (1997). It seems to me that "ought" claims and judgments involving other moral terms like "good," "bad," "right," and "wrong" are too closely linked to be separable in this way, but that is a topic I cannot fully address here.

Finally, there are others who do not state a clear argument, but who seem to take it as obvious that determinism would undermine moral ought claims. In the first *Critique*, using "nature" to mean a specifically deterministic conception of the world, Kant says: "indeed, the ought, if one has merely the course of nature before one's eyes, has no significance whatsoever." Kant (1781/1787) at A547/B575. According to Kant, "ought" claims do not apply to anything that is simply a part of a deterministic world. This claim is certainly too strong. We make perfectly sensible "ought" claims about all sorts of inanimate and apparently mechanistically determined objects (e.g., "The battery is charged, so the car ought to start"). The claim is a bit more plausible when restricted to what I describe below as "normative ought" claims, and I suspect that this is closer to what Kant had in mind.

avoided.[10]

In what sense are these conclusions unacceptable? I am confident not simply that the threatened moral claims are true, but that they are true *whether or not* determinism is true. Whether or not we live in a deterministic world, the Nazis should not have run concentration camps. And whether or not we live in a deterministic world, I should not have eaten so much last night. For all I know, the world is deterministic, and I can imagine learning that scientists and philosophers have settled with near certainty that the universe is deterministic. But I cannot imagine that this should lead me to give up these normative beliefs. Determinism or not, I simply cannot be persuaded that there is nothing I should have done differently.[11] As a result, it will not do to protect claims about what people should have done by simply insisting that determinism must be false. To defend these views, I believe it needs to be shown that the truth of determinism is not a threat to the "ought" claims we

---

[10] The argument from determinism sketched above also poses a significant problem for some of the claims made by those concerned about the relationship between determinism and moral responsibility. Compatibilists and incompatibilists about determinism and moral responsibility often have views about how we ought to treat those who violate apparent moral (and legal and societal) norms if determinism is true. For example, some incompatibilists are moved by the intuition that it is wrong to blame or to punish someone who could not have done otherwise. But if the argument above is correct, and if determinism is true, then the incompatibilist cannot truly say that we should not have imprisoned or executed criminals, or that we should not have blamed people. For if determinism is true, then we are or have been causally determined to blame and punish. We were unable to do anything else. And so, because "ought" implies "can," it would be false to say that we morally ought to do something other than blame and punish. There may well be other things that the incompatibilist might be able to say: that those people do not deserve blame or punishment, for example. But strangely this would not mean that we should have done anything differently. Therefore, the argument I have presented should be of considerable interest to those interested in the connection between determinism and moral responsibility.

[11] Admittedly, this may not be an argument that these claims must be true, but it is at least a good reason for seeking to defend and explain them.

normally take to be true.

In order to show this, we need to get clearer about the relationship between determinism and "can" and the relationship between "ought" and "can." As I shall argue here, there is a sense in which "ought" implies "can," and there is a sense in which determinism implies "cannot," but these are not the same senses of "can." This means that there is a reading of the argument from determinism on which this argument is valid, and there is a reading on which its premises are true. But there is no reading on which it is sound. In order to show this, I shall start with a quick clarification of the sense in which determinism makes it impossible for us to do otherwise. I shall then turn to a longer investigation of the sense in which "ought" implies "can."

III. How Determinism Limits What We Can Do:

Some may think that they can avoid the conclusion of the argument from determinism by rejecting the first premise, the idea that determinism makes it impossible for us to do anything other than what we actually do. So, let's briefly consider that first premise before turning our focus back to "ought" and "can."

According to the first premise, if determinism is true, then no person can do anything other than what she actually does. Though true, this claim is sure to have some compatibilists balking. Compatibilists, who believe that free action is compatible with determinism, often insist that even if a person were determined to act in a particular way, it may still be the case that she could do otherwise. Incompatibilists, on the other hand, often insist that determinism is

incompatible with freedom because if a person were determined to act in particular way, then she could not do otherwise. So, the compatibilist might complain that the argument from determinism depends on the controversial and unsupported assumption of an incompatibilist understanding of what it means to say that an agent can do otherwise.

This compatibilist objection should be set aside. The disagreement between compatibilists and incompatibilists is difficult to put to rest, in large part because both of their claims about the ability to do otherwise are true. If a person were determined to act in a particular way, then it is true that she cannot do otherwise, but it can also be true to say that she can do otherwise. While this may sound paradoxical, it is a simple and acceptable consequence of the fact that terms like "can" and "cannot" can be used with differing degrees of stringency. Thus, we can make the following, perfectly sensible statement: "Well, I can, and I can't." Compatibilists and incompatibilists make little progress by bickering about whether or not a person could do X in spite of the fact that she is determined to do not-X. To make progress, they need to settle the issue of which sense of "can do otherwise" is relevant to free action. Similarly, what we need to do is carefully identify the sense in which it is true to say that a person in a deterministic world cannot do anything other than what she actually does. Once we do this, compatibilists and incompatibilists alike should see that they can agree with premise 1 of the argument above.

The fact that "can" can be used with varying standards of stringency has been widely recognized. David Lewis's suggestion of a contextualist

account of "can" has had a lot to do with this. Lewis briefly describes and makes a case for his view in the following passage:

> To say that something can happen means that its happening is compossible with certain facts. *Which* facts? That is determined, but sometimes not determined well enough, by context. An ape can't speak a human language—say, Finnish—but I can. Facts about the anatomy and operation of the ape's larynx and nervous system are not compossible with his speaking Finnish. The corresponding facts about my larynx and nervous system are compossible with my speaking Finnish. But don't take me along to Helsinki as your interpreter: I can't speak Finnish. My speaking Finnish is compossible with the facts considered so far, but not with further facts about my lack of training. What I can do, relative to one set facts, I cannot do, relative to another, more inclusive set. Whenever the context leaves it open which facts are to count as relevant, it is possible to equivocate about whether I can speak Finnish.[12]

According to Lewis, then, when we say that an agent can do something we are saying that his doing that thing is compatible with a certain set of facts. The precise significance, and the truth value, of a claim that an agent can do something will depend crucially on something typically left implicit, the set of facts with which the agent's acting is being said to be compatible. Lewis's view is "contexualist" because he believes that the meaning of "can" on its own does not determine which set of facts is relevant, and instead, the relevant set is determined by the "context." So, for example, if we have been focused on discussing people's basic athletic abilities, it might be true to say that Jane can outrun Rita. But if someone takes note of the fact that Jane is wearing heavy

[12] Lewis (1976) p. 77.

boots and Rita is wearing track shoes, the context may make it false to say that Jane can outrun Rita.

We should separate two aspects of Lewis's view. First, Lewis suggests that to say that something can happen is to say that its happening is compatible with some set of facts and that *which* sorts of facts make up the relevant set may vary from one use of "can" to another. That is to say that different uses of "can" share a common semantic element; they all express the idea of compatibility with a set of facts. At the same time, this shared element does not provide the complete significance of a particular use of "can." The significance of "can" will vary from one use to another as the set of relevant facts varies. The second element of Lewis's view is that it is the "context" that determines the precise significance of "can." It does so by determining which set of facts is relevant. The first element of Lewis's view is that there is limited variation in the significance of "can." The second element is that it is context that determines the precise significance within that range.

It appears that these two elements of Lewis's view can be evaluated separately. One could believe that Lewis is right about the significance of "can" and its variability, while believing that he is wrong that context determines the precise significance of a particular use of "can." Of course, one could rule out this possibility by making the second element of Lewis's view a definitional truth, *i.e.*, by claiming that "contexts" consist of "whatever facts determine the precise significance of a use of the word 'can'." But this is not an entirely natural suggestion. Provided that one does not take this line,

then one might agree that “can” claims generally express compatibility with some set of relevant facts, but think that which set of relevant facts might be determined by something not naturally, or most usefully, described as the “context” of that use of “can.”

I point this out not to begin a discussion about the nature and plausibility of contextualism as a theory of what fills out the meaning of a given use of “can.” Instead, my goal is to avoid this issue. Discussion of what constitutes a “context” and how context determines the significance of a particular use of “can” would take us too far astray. My point here is that we may accept Lewis’s views that “can” expresses compatibility with a set of facts and that the relevant set of facts may be different from one utterance to the next, but that need not commit us to any particular view about how, in general, the relevant set of facts is fixed. A general theory of how the relevant set of facts is fixed is not necessary for our purposes. Thus, I will feel free to speak of their being different senses of “can,” without meaning to suggest that this is Lewis’s view and without meaning to commit myself to any technical theory about meaning or “senses.”

I take the first aspect of Lewis’s view about the meaning of “can” to be quite plausible. His own examples seem to me to do a good job of supporting his view, but let me add another. Suppose that a student is hired as a part time assistant to the administrators of a university department. Imagine that his job provides him with access to the student records and grades on the university database. So, it is possible for him to alter the grades, though it is

also fairly likely that if he does this he will be caught and subject to harsh sanctions. Suppose that he tells a friend about the opportunity that this job presents him, not because he actually wants to take advantage of it, but just because the idea of it is novel. Perhaps though, the friend asks him to change her grades. He might naturally say, "I can't do that." When she complains, "You just said you could," he might reply, "Well, I can, and I can't."

Lewis's description of the significance of "can" does an elegant job of explaining this response. The student can change the grades in the sense that his doing so is compatible with the facts about his access to the database. But the student cannot change the grades in the sense that doing so is incompatible with the facts about the consequences of getting caught, his interest in avoiding those consequences, and (we may hopefully add) his commitment to live up to the trust put in him and uphold academic integrity.

So, let us take it as established that "can" may be used with different, but very closely related senses. In general, as Lewis suggests, phrases like "A can X" can be understood as saying that A's X-ing is compatible with certain facts.[13] Which facts are relevant depends on the sense and stringency with which "can" is used. So, in which sense of "can" is it true that

[13] It is worth noting that this rough explanation of the meaning of "can" works not only when "can" is used to express what is possible, but also when it is used to express what is permissible. "Can" may be used to say that something is or is permissible according to some set of norms or rules, as in "You cannot treat people that way." When it is used in this way, "can" expresses the compatibility of a certain event with what is required by moral norms (or whichever set of norms is at issue). If one believes that these are not *facts* strictly speaking, then we might preserve the core of Lewis's suggestion about the semantics of "can" by replacing his use of "facts" with "facts and apparently fact-like things that make up the content of morality, etiquette, rationality, etc."

determinism limits what we can do?

If determinism is true, then a thorough description of the state of the world at any point prior to our actions, together with the laws of nature, entails a precise description of what we will do.[14] So, taking the distant past to be fixed, if determinism is true, then one would have to break the laws of nature in order to do anything other than what she will in fact do. So, if determinism is true, then one is unable to do anything but what one actually does in the sense that, given our history, doing anything other than what one actually does would be inconsistent with the laws of nature. In other words, if determinism is true, then it is *nomologically impossible* to do anything other than what one does, *given* the state of the world at any point prior to one's action.

It is important to see that the sense in which one cannot do anything but what one actually does is not that of metaphysical impossibility. It is weaker than metaphysical impossibility in two ways. First, what I am calling "nomological impossibility" may be different than metaphysical impossibility. A

---

[14] There are, of course, some sticky issues that I am passing over quite quickly here. A thorough discussion would need to tell us more about what counts as a "thorough description of the state of the world at a time." We mean to say that the arrangement of the world at a given time together with the laws of nature determine what will happen in the future. They determine what will happen through the progression of causally linked events. But if our description of the state of the world at a given time is too inclusive it may trivially entail what will happen in the future. For example, if a description of the world at time includes descriptions of the temporal relations of objects to future objects and events (*e.g.*, "At time *t*, there is a dog at location *l*, that is one minute away from falling asleep.") then the descriptions of the state of the world will, by themselves, entail what will happen at later times. We must suppose that the descriptions of the state of the world do not include these temporal relations. Saying more specifically what these descriptions of the state of the world would be like is a task that would take us too far astray. For some further discussion of this issue, see Van Inwagen (1983) pp.59-60.

state of affairs or event occurrence is nomologically impossible, if it is not a constituent of any possible world with the same natural laws as the actual world. So, in considering whether something is nomologically possible or impossible by thinking in terms of possible worlds, we consider only the possible worlds with the same laws of nature. Those worlds may or may not make up only a proper subset of all of the possible worlds, depending on whether or not there are possible worlds with different laws.

Second, the truth of determinism makes certain actions nomologically impossible only relative to the causally relevant features of our history. Determinism limits a person to what she actually does through a combination of the laws of nature and the relevant background conditions. For example, it is not nomologically impossible for me to long jump 20 feet. It seems that there is a possible world with our laws in which I jump 20 feet. That world is one in which I have developed a good deal more of the right sorts of muscle than I actually have. So, determinism would not threaten the nomological possibility of my jumping 20 feet. Nonetheless, if determinism is true it may be nomologically impossible for me to jump 20 feet in any world with a history quite like our own. In any actual situation in which I might attempt a long jump, my actual speed, strength, and size will be causally relevant to the results of my attempt. Given the history that has led to these causally relevant features and given gravity as it actually is, if I attempt to jump, I will not jump 20 feet.

So, when we say that determinism makes it impossible for me to do anything other than what I actually do, or when we say that I cannot, because

of determinism, do otherwise, we are saying that my doing otherwise is incompatible with the laws of nature taken together with all of the causally relevant facts about the past. The sense of "cannot," then, in the first premise of the argument is a sense expressing this sort of incompatibility. The argument would involve no equivocation and would clearly be valid if the sense of "can" in "'ought' implies 'can'" expresses compatibility with this same set of facts. That is, the argument would be valid if premise two says that if a person ought to do something in a given situation, then she can do it in the sense that performing that action is compatible with the laws of nature and the causally relevant features of the past. As we turn to an examination of the relationship between "ought" and "can," we will see whether any such premise is plausible.

IV. Some Objections to "'Ought' Implies 'Can'":

In what sense, if any, does "ought" imply "can"? We will need to determine which uses of "ought" imply which uses of "can," and we will want to know why. Let us start that investigation by looking at a criticism of the idea that there is a logical or conceptual relationship between "ought" claims and "can" claims. If it were successful, this criticism would allow us to avoid the argument from determinism by rejecting Premise 2.

Walter Sinnott-Armstrong has argued that "if 'ought' entailed 'can', an agent could escape having to do something simply by making himself unable

to do it."[15] Suppose that I could do 20 pull-ups, *if* I trained steadily for 3 months. Suppose also that through some strange set of circumstances, it becomes the case that I ought to do 20 pull-ups on date D (3 months from today). This "ought" claim is compatible with any common sense understanding of "'ought' implies 'can'." But if "ought" does imply "can," then it looks like it becomes false that I ought to do 20 pull-ups on D at the moment I no longer have enough time to get in shape. If I simply decide I prefer a regimen of fatty snacks, reality TV, and heavy drinking, I will not be able to do 20 pull-ups when the time comes. In that case, it would literally take a miracle for my puny arms to lift my flabby body off the ground. So, the supporter of "'ought' implies 'can'" seems committed to saying, at the moment I stray too far from my training regimen, that it is no longer the case that I ought to do the pull-ups.

One worry that this sort of case brings out is that this would make it too easy for us to get out of what we ought to do. Sinnott-Armstrong worries that I can escape my duty by simply getting myself into a position where I could not possibly get into shape in time.[16] I agree that this consequence would be intolerable. Surely one cannot "escape" a duty by purposefully, or negligently, making oneself unable to fulfill it. If this were actually a consequence, then any moral philosopher appealing to "'ought' implies 'can'" would have made a terrible mistake. But "'ought' implies 'can'" does not have this consequence.

---

[15] Sinnott-Armstrong (1984) p.252.

[16] *Id.*

If “ought” implies “can,” it may well be false that I ought to X once I get into a position in which I cannot X, but that hardly shows I have “escaped” my duties in any morally significant sense.

To see why this is, consider Sinnott-Armstrong’s example:

> Suppose Adams promises at noon to meet Brown at 6:00 p.m. but then goes to a movie at 5:00 p.m. Adams knows that, if he goes to the movie, he will not be able to meet Brown on time. But he goes anyway, simply because he wants to see the movie. The theater is 65 minutes away from the meeting place, so by 5:00 it is too late for Adams to keep his promise. Consequently, if ‘ought’ entailed ‘can’, it would not be true at 5:00 that Adams ought to meet Brown. Similarly, if Adams is still at the theater at 6:00, he cannot then meet Brown on time. Consequently, if ‘ought’ entailed ‘can’, it would not be true at 6:00 that Adams ought to meet Brown.[17]

Everything Sinnott-Armstrong says here is correct, but it is also unproblematic for the idea that “ought” implies “can.” It does not show that Adams has somehow “escaped” his obligation, at least, not if “escape” means to have removed the obligation and responsibility for failing to meet it. In the example, Adams had promised to meet Brown, and at that time he ought to fulfill his promise. But in order to satisfy his whim, Adams makes it the case that he cannot meet Brown. Suppose that “ought” does imply “can.” Once it is 5 p.m. and Adams can no longer fulfill his obligation to meet Brown, the claim that he ought to meet Brown becomes false. This fact would be odd if it meant that Adams had somehow relieved himself of responsibility. But nothing of the

---

[17] *Id.*

sort follows. It is unstated in the example, but certainly true, that Adams ought not go to the theater. So, it is still perfectly true, at 5 p.m., that Adams *ought to have* done what was compatible with meeting Brown, and at 6 p.m., it is also perfectly true that Adams *ought to have* met Brown. So, we should maintain throughout that Adams had an obligation that he failed to satisfy, and that he is, all other things being equal, blameworthy for this failure. In short, holding that "ought" implies "can" does not force one to think that a person can, in any significant sense, "escape" obligations simply by making himself unable to perform them.

But we may not yet have avoided what is most problematic about cases of self-imposed impossibility. Sinnott-Armstrong has a linguistic argument, as well as a moral argument. As Sinnott-Armstrong goes on to point out, someone holding that "ought" implies "can" may have to deny a number of very intuitive "ought" claims that could be made in a case like the one above. He writes:

> If Adams calls Brown from the theater at 6:00, it would be natural for Brown to say, "Where are you? You ought to be here (by now)," even though Brown knows that Adams cannot be there. Brown's statement seems true, because Adams did promise, the appointment was never mutually cancelled, and the obligation is not overridden. Thus, there is no reason to deny Brown's statement except to save the claim that 'ought' entails 'can', and that reason would beg the question. Furthermore, if Adams calls at 5:00 and tells Brown that he is at the theater, Brown might respond, "Why haven't you left yet? You ought to meet me in an hour, and it takes more than an hour to get here from the theater." Again, Brown's

statement seems natural and true.[18]

In this passage, Sinnott-Armstrong shows that it is sometimes natural to say that something ought to be the case even though it has become impossible for a person to make that thing happen. Adams ought to be at his meeting as promised, but it has become impossible for him to be there. So, Sinnott-Armstrong feels that he has a counterexample to the idea that "ought" implies "can."

People often object to the idea that "ought" implies "can" by finding some "ought" claim that seems to be true and showing that if we replace "ought" with "can" we get an intuitively false claim. They suppose that if "ought" implies "can", than every claim involving the word "ought" implies a quit similar claim with the "ought" replaced by "can." As we have just seen, it is not difficult to come up with counterexamples if one takes this broad formulation of "'ought' implies 'can'." But we have not yet committed ourselves to any specification of the general idea that "ought" implies "can." So, when confronted with these counterexamples, we must ask whether they show that we were wrong to think that "ought" implies "can" or, instead, that "'ought' implies 'can'" should be more carefully stated as the principle that certain kinds of "ought" claims imply certain kinds of "can" claims. Here at least, I believe that the latter is the case.

Let me make clear what I have in mind. I have just suggested one not too subtle understanding of the phrase "'ought' implies 'can.'" On this

[18] *Id.*

understanding, which I shall call the Broad Formulation, the phrase is taken to mean that any sentence using the term "ought" as a predicate implies a similar sentence with the word "can" replacing the word "ought."[19] So, from "I ought to keep working," we may infer "I can keep working." From "I ought to be on a sandy beach with a drink in my hand," we may infer that "I can be on a sandy beach with a drink in my hand." And also, from "there ought to be a law against that," we may infer "there can be a law against that." Contrast this Broad Formulation with the second premise of the argument from determinism. That premise reads: If a person ought *to do* something, then he or she can do that thing. Premise 2 differs from the Broad Formulation because it concerns "ought" only when it is used to connect agents and ways of acting. It concerns the implications of saying that an *agent* ought *to do something*.

There are all sorts of uses of "ought" that do not say that an agent ought to do something: there ought to be food on every plate; there ought to be more love in the world; you ought to be home by now. These "ought" claims are relevant to the issue of how agents ought to act, but typically it takes more information to get from these claims to any informative claims concerning the courses of action that agents ought to pursue. There ought to be more love in the world, but that does not entail that I ought to do anything in particular to bring about more love. Perhaps there are other things that it is

---

[19] To reiterate: According to the Broad Formulation, "Blah blah *ought* yada yada" entails "Blah blah *can* yada yada" no matter what we substitute for "blah blah" and "yada yada."

more important for me to focus on. Or it may be that I ought to be somewhere else, supporting a friend say, but if I cannot get there in time, nothing follows about what I ought to do. Perhaps I should try to be there as soon as I can. Perhaps I should simply stay put and call instead. So, though Premise 2 expresses a relationship between some "ought" claims and some "can" claims, it does not commit us to a view about the implications of all "ought" claims.[20]

We have two formulations of the principle that "ought" implies "can," the Broad Formulation and Premise 2. Of course, we might come up with others, but these two are enough to allow us to evaluate Sinnott-Armstrong's counterexamples. There are two "ought" claims that Sinnott-Armstrong sees as problematic here: "You [*i.e.*, Adams] ought to be here (by now)," and "You [*i.e.*, Adams] ought to meet me [*i.e.*, Brown] in an hour." Each of these claims strikes Sinnott-Armstrong as natural and true, though Adams cannot be there and cannot meet Brown (on a common sense interpretation of "can"). Indeed, Sinnott-Armstrong suggests that the only reason we would have for denying these "ought" claims is the question-begging reason of trying to uphold the idea that "ought" entails "can."[21]

First, consider the claim that Adams ought to be there, at the agreed meeting place, now. This claim does strike me as undeniable. He promised to be there. Nothing overrode or dissolved this commitment. So, he ought to

---

[20] The relationship between claims about what ought to be the case and claims about what agents ought to do is discussed in greater depth in the next section of this Chapter.

[21] Sinnott-Armstrong (1984) p.252

be there. In addition, let us grant that it is also true, at the same time, that he cannot be there. While we should grant all of this, we should also be clear that nothing follows about what Adams should do now. Perhaps he ought to make an attempt to get there as soon as possible. But if Brown is not willing or able to wait, then this attempt would be pointless. So, it is just as likely that he ought to apologize and set up a new meeting time. What he ought to do depends on facts that Sinnott-Armstrong never even considers. In any case, whatever it is that he ought to do, we have no reason to think that it is something he cannot do. So, while Sinnott-Armstrong might have identified a counterexample to the Broad Formulation, it is not a counterexample to Premise 2, the principle that if an agent *ought to do something*, then he or she can do that thing.

Consider the second problematic claim, the claim that Adams ought to meet Brown. This one is not so clear-cut. Meeting someone is something that an agent might do. "I am meeting a friend" is a perfectly good answer to the question, "What are you going to do tonight?" So, if Sinnott-Armstrong is right about the claim that Adams ought to meet Brown, then we may have a counterexample not just to the Broad Formulation, but also to the narrower Premise 2. However, it is not so clear that Sinnott-Armstrong is right. Premise 2 is a principle concerning what an agent ought to do, a principle concerning which courses of action he ought to pursue. It concerns claims that are potential answers to the question, "What should this agent do?" While I do not find it absurd or unnatural to say that Adams ought to meet Brown in the

conversation that Sinnott-Armstrong describes, I do find it to be an inappropriate answer to the question, “What should Adams do?” As noted above, it looks like Adams should go meet Brown, only if, among other things, doing so is still possible for Adams and desirable for Brown. Otherwise, going to the meeting place would be a waste of time, and Adams ought to apologize and the two should determine how best to proceed. So, as an answer to the question, “What should Adams do?”, the truth of the suggestion that Adams ought to meet Brown does depend on whether or not Adams can meet Brown.

I am conceding that it might be conversationally acceptable to say that Adams ought to meet Brown, even though Adams cannot meet Brown. But I insist that that statement is false if it is meant to say what course of action Adams ought to pursue. It might be conversationally acceptable simply because it is close enough to expressing a true statement like “You [Adams] *ought to be meeting* me [Brown]” or “You were supposed to meet me” or, of course, “You ought to have met me.” These claims all can be true without violating Premise 2, even though it has become impossible for Adams to make it to his meeting.

In sum, Sinnott-Armstrong’s examples do not convince. This relationship between “ought” and “can” does not make it possible for agents to escape their duties. And the counterexamples we have looked at do not work. One is unconvincing in general, and the other is a counterexample only to the Broad Formulation of the principle.

Limiting the principle to “ought to do” claims also disarms a separate

objection: that we often ought to feel a certain way or ought to believe something even if we cannot. We do not have the same control over our feelings that we suppose we have over our actions. We may feel jealous of a person even though we know we should like him. In the short term, we may not be able to prevent these feelings of jealousy from arising even though we should not feel them. We might also find it difficult to rid ourselves of beliefs or prejudices even after we have come to recognize that we should not hold them. These are not counterexamples to the principle that if a person ought to do something, then the person can do that thing. If we feel something that we should not feel, or if we hold a belief that we should not hold, then we ought to do what we can to change those feelings or beliefs. This may mean not acting on those beliefs or feelings, and it may mean taking steps that will gradually result in change. None of this is inconsistent with the principle that we ought to do only what we can.

Of course, it is one thing to note that we can avoid counterexamples to "'ought' implies 'can'" by adopting a limited version of the principle. It is another thing to explain why this more limited version makes sense. Sinnott-Armstrong's counterexample does provide us with a challenge. We must be able to provide a clear account of the distinction between different kinds of "ought" claims, and we must be able to explain why one kind implies "can," while others do not.

V. The Many Ways of Using "Ought":

This section takes up the challenge of explaining which "ought" claims

imply “can” and why. A number of philosophers have attempted to distinguish the uses of “ought” that imply “can” from the uses that do not by identifying different “senses” of “ought.” If “ought” has what philosophers like to call different “senses,” then this may provide the explanation of why different “ought” claims have different logical relations to claims about what is possible.

But the suggestion that “ought” has distinct senses can lead us to underemphasize the common semantic contribution made by all uses of “ought.” In general, we use “ought” to say that an event, action or state of affairs is favored or supported by some set of relevant facts and considerations. Earlier, we saw that, though the significance of saying that something *can* be the case may vary, “can” has a common core meaning or function across its various uses. The same is true of “ought.” When we use “ought,” we make reference (often implicitly) to some set of considerations, and we say that those considerations support the event, action or state of affairs that we are saying ought to be. This is true whether we are using “ought” to make a claim about how we expect things to actually happen, as in, “Because he left an hour ago, John ought to be home by now,” or to make a claim about how it would be good for things to happen, as in, “Because he promised, John ought to be home by now.” In both cases, we suggest that John’s being home by now is supported or favored by some set of relevant considerations.

A. Predictive Oughts and Normative Oughts

While suggesting that there is a common, core meaning to the term

"ought," I also suggested in passing that there is a distinction between two broad kinds of claims involving "ought." Allow me to clarify this distinction. We have theories about how the world or suitably small parts of it tend to work and to behave. These may be theories of the grand scientific sort, or they may just be groups of pretty mundane beliefs. For example, I believe that my coffee maker, a filter, coffee grounds, and water combined in a certain way reliably lead to a fresh pot of coffee. (In my house at least, this is the extent of much of the theorizing that happens early in the morning.) We also have theories about the way the world, or some suitably small part of it, is arranged at a particular time. For example, I believe that things are combined in the coffee-making way, *i.e.*, the grounds are in the filter, which is in the coffee maker; water is in the water chamber; and the power is on. With some of our uses of "ought," we refer implicitly to these two sorts of considerations, as in: "There ought to be a fresh pot of coffee soon." A statement like this is typically meant to indicate that our theories about the arrangement and behavior of this part of the world support the presence of coffee in the pot.

The connection to our theories regarding the arrangement and behavior of the world is not always left implicit. We tend to make the connection explicit when things do not go as we expect: "I put the grounds and the water in there, and I turned it on. So, there ought to be coffee. Why isn't this working?" We might also make the connection explicit when we are testing the relevant theory or when we are not yet confident in the theory. For example, biologists may say things like: "According to our model, we ought to

find more fish in the colder, shade-covered sections of the stream."

When "ought" is used to express the support of theories intended to capture how things actually are and how they actually tend to behave, it is used to express a either a prediction or an expectation based on empirical evidence. One mark of this "predictive ought" is that when things do not turn out to be as we say they ought to be, this gives us a reason to believe that our theories are incomplete or incorrect. The lack of coffee is a reason for me to suppose that things are not arranged as I thought they were or that there is some other relevant factor that my theory of coffee maker behavior overlooks.[22] Or if the biologists just mentioned fail to find more fish in the colder parts of the stream, they may scrap their theory altogether or decide that it needs to be modified to incorporate some additional variable that also affects the distribution of fish populations.

Of course, many "ought" claims are not like this. Many "ought" claims are based not on theories that are intended to capture the actual behavior of our world, but on theories about what is good or what is right. They are based on our theories about what is moral, rational, or polite, for example. They can also be based on considerations about what is legal or what would comply with the rules and goals of a game. Suppose that I am expecting guests, and I

[22] It is possible that my expectations of coffee might have a normative element as opposed to a predictive element. I might, for example, feel that there ought to be coffee because Sears and/or Mr. Coffee should have sold me a *good* coffee maker, not one that is defective or difficult to operate. I do not think this must be so, but to avoid confusion, assume that I found the coffee maker used and discarded, and that I only believe that it should work because, looking it over, I cannot find any significant problem with it.

am thinking about what I will need to do to be a good host. I might say to myself, “I ought to provide coffee” or “There ought to be coffee in the morning.” I am not making a prediction. I am not even saying that there is any basis for such a prediction. After all, I might go on: “But there’s no chance I’m going to the store to buy beans tonight.” All that I am saying is that it would be good, considerate, or polite of me to provide coffee. These “ought” claims express the support of my theories about what guests enjoy and about what good hosts do. Good hosts try to meet the reasonable demands of their guests, and coffee is a likely and reasonable demand of many guests. So, I conclude that I ought to have coffee available.

Of course, when we use “ought” to express the support of a theory about what would be right, good or appropriate, we often find that the world does not conform. People act immorally and irrationally; people break laws and flout the rules of games. In contrast to the “predictive ought” we do not take this non-conformity to suggest that the relevant theories are wrong or incomplete. Consider an example. Suppose that the wealthy ought to do more than they actually do to help the poor. It follows trivially from this that things are not as they ought to be. With the predictive use of “ought,” when things are not actually as they ought to be this is a reason to suspect that our theories are wrong, but in this case, it is a reason to suppose that people are behaving wrongly. For lack of a better term, I call uses of “ought” that express the support of theories about how it would be good or appropriate for things to be “normative” uses of “ought.”

When people suggest that "ought" implies "can," they do not typically have the predictive "ought" in mind. "'Ought' implies 'can'" is typically taken to be a principle about either normative "oughts" in general or some sub-category thereof. But I see no reason to suppose that there is no entailment relationship between the predictive "ought" and "can." In fact, it seems that there is a pretty obvious relationship. When we use "ought" predictively, we suggest that certain considerations provide support for the actuality of a particular state of affairs. Of course, if these considerations support that state of affairs, then they are also compatible with it. The facts and considerations that support the existence of coffee in the pot are also compatible with existence of coffee in the pot. The "can" claim seems only to be weaker than the predictive "ought" claim, and therefore it is entailed by the "ought" claim.[23]

Nonetheless, when people suggest that "ought" implies "can," they do

---

[23] Of course, there are cases in which we want to say that there ought to be coffee in the pot, in the sense that we would expect coffee to be there, but nonetheless, there is no coffee there. In this case, we presume that there is an explanation for the lack of coffee, and so, we presume that there is some fact that is incompatible with the presence of coffee in the pot. This fact causally explains the absence of coffee. Once we accept that there is such a fact, we can say that there cannot be coffee in the pot; something is preventing it. In this case, it is both true that there ought to be coffee in the pot and that there cannot be coffee in the pot. This can be explained without giving up on the idea that the predictive "ought" implies "can." In this case, the set of facts relevant to the "ought" claim differs from the set of facts relevant to our "can" claim. When we say that there ought to be coffee in the pot (but there is not), we have good reason to believe that our coffee-making theory fails to capture all of the coffee-relevant considerations. Nonetheless, we can say that there ought to be coffee in the pot if we mean, "Given what I know about coffee making and the way things are in this kitchen (*i.e.*, given my coffee making theories), there ought to be coffee in that pot." Once we insist that there ought to be coffee, even though there is not, we are implicitly restricting the support for the existence of coffee to our admittedly incomplete theory and not to a general and accurate description of the facts. If we keep the focus restricted to that set of considerations, then that use of "ought" surely does imply "can." The incomplete theory of coffee making supports the presence of coffee. So, surely it is also compatible with the presence of coffee. When we say that there cannot be coffee in the pot, something is preventing it from being there, we are taking into consideration some facts and considerations that are left out by our incomplete coffee-making theory.

not generally have in mind the rather boring thought that if a set of empirical considerations supports something, then that set is compatible with that thing. Instead, they have in mind some form of the rather interesting thought that if a thing is supported by a theory of what is right or what is good or what is appropriate, then that thing is compatible with the way the world actually is and the way it actually operates. In other words, they mean that a normative "ought" implies a predictive "can." So, let us now focus our attention on the normative "oughts."

B. A Supposed Distinction Between Moral and Other Normative Oughts

It is common to distinguish different normative "oughts" based on the kinds of evaluative or normative considerations that are salient. Many philosophers distinguish between the "oughts" of morality, prudence and etiquette, for example. The "moral ought" is used when moral considerations are relevant, the "prudential ought" is used when prudential considerations are relevant, and so on. While drawing such distinctions may be helpful for some purposes, I shall not do so for a couple of reasons. First, I can give no good account of the differences between these different sorts of considerations. Surely, we can see that murder is morally wrong, not merely impolite and unwise. Similarly, whether or not to R.S.V.P. does not usually bring up a moral issue. But between these extremes there is a whole set of issues that defy easy categorization. We may cause offense to people in all sorts of ways, some of which are merely impolite, others of which are quite seriously

disrespectful. That there would be a clear line dividing the immoral from the impolite, rather than a spectrum connecting them, seems implausible to me. And of course, the idea of a clear line between the moral and the prudential, and between the moral and the rational, is also challenged by a range of rich ethical theories.[24]

More importantly, whether an "ought" claim is based on considerations of morality, rationality, etiquette, or what have you, does not seem to have a significant bearing on that claim's relations to claims about what is possible. The "'ought'" implies "'can'" principle is usually discussed as a principle of moral philosophy, but this alone is no reason to suppose that its applies only to true moral issues. The example that began this Chapter, involving Sue the veterinarian, did not concern a moral issue, but it is as clear an example of "ought" implying "can" as any I can think of. If we want to know what a person ought to do, *i.e.*, which actions the balance of evaluative or normative considerations weighs in favor of, then it seems that what that person can do will be relevant in the same ways, whether the most salient reasons are moral, prudential, or whatever else. So, we should at least start from the assumption that, whatever differences there are between moral, rational, and prudential considerations, these differences do not affect the relationship between "ought" and "can."

---

[24] I have in mind here broadly Aristotelian theories according to which being virtuous is part of a flourishing or truly happy human life; rational egoism or Hobbesian theories according to which moral and political duties are shown to be a requirement of enlightened self-interest; and broadly Kantian theories that suggest that morality is a certain kind of rationality.

C. *Prima Facie* Oughts and All Things Considered Oughts

One note of caution is needed here. When discussing the "predictive ought" we saw that we may qualify an "ought" claim by adding (implicitly or explicitly) something like "given this model of how things behave" or "given what we know about this topic." When we qualify a predictive use of "ought" in this way, we do not imply that what ought to happen actually *can* happen, *all things considered*. We acknowledge that there may be some important, relevant considerations that will prevent things from being as we would expect them to be. The same thing may occur with normative "ought" claims. We may say, "According to the rulebook, the referees ought to confer and decide who had the best perspective on the play," or, "Legally speaking, you should not have done that." When we (implicitly or explicitly) make these sorts of qualifications, we are acknowledging that other factors may be relevant to what "ought," in the normative sense, to be the case, all things considered. It may be that, *according to the rulebook*, the referees ought to confer, but if the coliseum is collapsing, then the referees ought, *all things considered*, to get themselves to a safe place, not waste their time conferring about the play. Now, when we qualify a normative "ought" claim in the way that I am suggesting, that is, when we implicitly or explicitly suggest that we are not focusing on all of the relevant normative considerations, then our "ought" claim may not imply "can." Going back to the referees' rulebook, it quite likely makes no exception for cases in which the referees are bound, gagged and thereby unable to confer. So, at least on one strict reading, it may be true that

according to the rulebook, the referees ought to have conferred even though it was impossible for them to confer.

This is not a serious problem for the idea that "ought" implies "can." When we qualify a normative "ought" claim by saying that we are taking only a limited range of normative considerations into effect, we are often saying, "According to this limited set of considerations, this is what a person ought to do, *but* this may not really be what you ought to do." We may also think of these qualified "ought" claims as implicitly conditional in form. "According to considerations 1, 2, and 3, agent A ought to X" means "If 1, 2, and 3 were the only relevant considerations, then agent A ought to X." Of course, conditionals like this may be true while it is false that A ought to X. Once qualified "ought" claims are understood in this way, then it is not surprising that they do not imply "can."

In what follows, I shall focus on "all things considered" normative "ought" claims. When we mean to suggest that the balance of all relevant normative considerations supports a certain event, action or state of affairs, then it seems to me to make no difference whether the salient considerations include the law, morality, prudence, grammar, or anything else of the kind. If, *all things considered*, a person ought to do something, then it seems intuitive that it must be the case that this person can do that thing.

D. The Ought to Do and the Ought to Be

Now we are prepared to turn to the distinction that I relied on in the previous section: the distinction between claims about how agents ought to

act and claims about how things ought to be. I have suggested that it seemed intuitive that if an agent ought to do something then that agent can do that thing, but that it did not seem intuitive that if things ought to be a certain way, then things can be that way. My aim now is to give a principled explanation of this intuition.

First, let us rule out a possible misconception. To say that there is a distinction between "ought to do" and "ought to be" could be taken to mean that what matters is the wording of such claims. It could suggest a view along the following lines. Though an "ought to do" claim rarely involves the precise phrase "ought to do," it will contain the phrase "ought to x" where x is replaced by some verb connoting action. An "ought to be" claim, on the other hand, contains the phrase "ought to x" where x is replaced by some tense of the verb "to be," or perhaps "to have" or some other suitable verb. This will not do. This difference in wording does not track any real distinction. One can express precisely the same idea with phrases on either side of the distinction just described. Consider "John ought to tell her the truth about X" and "John ought to be honest with her about X"; or "The most worthy candidate ought to win" and "The most worthy candidate ought to be victorious"; or "Sonia ought to judge your case impartially" and "Sonia ought to be impartial in judging your case." On the suggested account of the distinction the first of each pair is an "ought to do" claim, and the second of each pair is an "ought to be" claim. But this is surely wrong. If we want to think of the distinction as one between "ought to be" and "ought to do," we should not be led into thinking it is a

distinction that can be tracked by the wording of "ought" claims. Instead, we should take the distinction between the senses to coincide with the distinction between saying, by whatever words, that an agent ought to pursue a course of action and saying that the world ought to be a certain way.

Others have suggested that some such distinction is relevant to the "ought" implies "can" principle, and some have at least gestured at rationales for this distinction. For example, G.E. Moore said that the distinction between "ought to do" and "ought to be" is closely related to a distinction between "rules of duty," on the one hand, and "ideal" moral rules, on the other. [25] He claims that the "ought to do" sense of "ought" is used to express a "rule of duty." It is used to tell us what people have a duty to do. The "ought to be" sense of "ought" is used to express something more "ideal." It tells us how it would be good to be, ideally speaking. Such ideal rules and the related "ought" claims do not imply that we have a duty to do something in particular. Instead, they simply say what the ideal is. It appears that the fact that an "ought to do" expresses a "rule of duty" is supposed to figure into the explanation of why an "ought to do" implies "can," though Moore does not explain precisely how.

Similarly, Michael Zimmerman has claimed that the "ought to do"/"ought to be" distinction is best explained in terms of a "binding" and a "non-binding" sense of "ought."[26] "Ought", he has said, is used in its "binding" sense when it expresses an obligation. It is used in its "non-binding" sense when it

---

[25] Moore (1922) pp.315-22.

[26] Zimmerman (1996) pp.2-5.

expresses the idea that something would be good or ideal, but not that someone has an obligation.

These suggestions do not explain the differing implications of the "ought to do" and the "ought to be." I am interested in distinguishing the "ought" claims to which the "'ought' implies 'can'" principle does apply from those to which it does not apply. But that distinction is not the supposed distinction between "duties" and "ideals" or between "binding" and "non-binding" uses of "ought." Suppose that Maude is generally a very good-natured person. However, she has certain pet peeves that can put her into a somewhat foul mood. They do not make her especially mean or nasty, only a bit less kind and a bit more easily annoyed than normal. On the infrequent occasions that Maude is affected by her pet peeves, she tends to do a good job of not taking this out on undeserving people. So, on the whole, people who know Maude thinks she is a wonderful person, and many of them even strive to emulate her. Perhaps, though, there are certain steps that Maude could take to overcome her pet peeves and to become even more consistently kind. Suppose that a therapist could help her to be less annoyed by them. Now, I think that it is plausible that Maude morally ought to go see the therapist since this would help her to be a more consistently virtuous person, but I suppose also that she has no strict obligation or duty to do so. After all, she is already an exceptionally good person, and the therapy, let us suppose, would be rather long and difficult work. So, Maude ought to see the therapist, if she is to approach the moral ideal, but she is not blameworthy if she does not. Now, it

is not exactly clear what Moore means by the "ideal" sense of "ought" or what Zimmerman means by the "non-binding" sense, and that in itself is a problem. But it seems to me that the claim that Maude ought to see the therapist is plausibly described as "ideal" and "non-binding." This is a problem. First, it looks like "Maude ought to see a therapist" is, on the one hand, "ideal" and "non-binding," and on the other hand, a claim about what Maude "*ought to do*." Thus, it is not clear that the "ideal"/"rule of duty" distinction, whatever that distinction is, lines up with any intuitive notion of the "ought to be"/"ought to do" distinction.

Furthermore, the "ideal" claim that Maude ought to see a therapist does imply that Maude can see the therapist, and it seems to do so for quite the same reasons that a binding "ought" claim would imply "can." After all, suppose that Dr. Brown, the only therapist capable of the appropriate therapeutic techniques, suddenly retires and refuses to see any more patients. At this point, it becomes false to suggest that Maude ought to see Dr. Brown (or any other therapist) for treatment. She cannot see Dr. Brown, or anyone else, for treatment. My intuitions about this do not change whether we take the case as it is or whether we imagine that Maude's problem is more serious, giving her a much more serious and binding reason to see the therapist. Given that she cannot see a therapist, it makes little sense to suggest that this is what she ought to do. So, the suggestion that the "ought to do" is "binding" or related to "rules of duty" does not seem to explain why the "ought to do" would imply "can," and the suggestion that the "ought to be" is "non-binding" or

"ideal" does not explain why it does not imply "can."

Maude's case might appear to be contrived, but I suspect that something analogous is extremely common. We often feel that we ought do things that we are not strictly obligated to do. For example, we should support our family and friends by attending their significant life events, their performances, their parties, etc. For any such event, it may be true that I ought to attend, but that I am not bound or obligated to do so. I may miss some of these events, sometimes because of conflicting obligations, but sometimes simply because I am tired and want some time to myself. Nonetheless, when I truly cannot attend a particular event, it would be bizarre to say that I ought to attend. Therefore, the supposed distinction between "ideal" and "binding" uses of "ought" does not explain a distinction between "ought to be" and "ought to do."

Gilbert Harman offers a suggestion that provides a better start to explaining the difference between the "ought to do" and the "ought to be." In *The Nature of Morality*, Harman notes that an "ought to do" implies that an agent has a reason to perform an action, whereas an "ought to be" evaluates a state of affairs and does not by itself imply that any particular agent has a reason to contribute to bringing about that state of affairs.[27] This seems to me to be a first step towards explaining why "ought to do" implies "can," while "ought to be" does not. It is intuitive that one does not have reason to perform actions that one cannot perform. So, the claim that an agent ought to do

[27] Harman (1977) pp.84-87.

something expresses the fact that she has a reason to do that thing, and this in turn implies that she can do that thing. In contrast, we can make sense of evaluating states of affairs that are, for all practical purposes, impossible. For example, we do say that it would be good if I could be in such and such place now, even if I cannot be there now.

However, things are a bit more complicated than this. It is not as though "ought to be" claims bear no relation to an agent's reasons for action. Claims about what ought to be the case do imply that agents have reasons of some sort. When we say that something ought to be the case, we say that normative considerations speak in favor of that thing's being the case. In other words, we say that it has a certain sort of value. If a state of affairs has value according to the balance of normative considerations, then any agent to whom those normative considerations apply has some reason to act in ways that respect the value of that state of affairs.[28] Often this will make little difference to a person's actual behavior because, for many valuable states of affairs, there will be little that the agent could do that would either show respect or disrespect for those states of affairs. But still, the agent has at least

[28] The phrase "any agent to whom those normative considerations apply" is meant to leave open the question of who ought to be moral, rational, etc. For some sets of normative considerations, it may be up to the agent whether or not they are binding on him. In the Kantian terminology that some have adopted, these normative considerations may underwrite only "hypothetical imperatives." To take an example, though a doubtful one, the fact that something would be polite might only be relevant to agents who care about being polite. Additionally, an agent's nature may make it the case that only some normative considerations can apply to it. There are things that non-human animals ought to do, in the normative sense. We might quite reasonably say that wild animals ought to do specific things that will help them to survive, and that working animals (*e.g.*, herding dogs, work horses, pack mules, etc.) ought to perform particular tasks. But because as far as we know most non-human animals cannot grasp and apply ethical principles, moral considerations do not apply to them.

some reason to avoid behaviors that would fail to respect the value of a state of affairs that ought to be the case.

So, in order to be careful about the relation between normative "ought" claims and reasons for action, we should note that there is a close connection not only between reasons and claims about what agents ought to do, but also between reasons and claims about what ought to be the case. Having noted this though, we may still maintain that the two relationships are different: If an agent *ought to do* something, X, then that agent has a reason to do that thing, X, but if something, Y, *ought to be* the case, then agents to whom the relevant considerations apply have a reason to act in ways that respect the value of Y. Does this difference permit an explanation of why only the "ought to do" implies "can"? I believe it does.

I suggested above that the relationship between normative "ought" claims and reasons for action might explain why the Broad Interpretation is false, while Premise 2 is true. It would help to explain this on the assumption that an agent has a reason to do something only if she can do that thing. Let us grant this assumption for a moment. If an agent ought to do something, then that agent has a reason to do that thing, and this in turn implies that that agent can do that thing. On the other hand, if something, Y, ought to be the case, then agents to whom the relevant considerations apply have a reason to act in ways that respect the value of Y, and this in turn implies that those agents can act in ways that respect the value of Y. Notice though, agents might be able to act in ways that respect the value of a certain state of affairs

even though that state of affairs is, in some important sense, impossible. Recall Adams, who can no longer make his meeting with Brown. Even after it has become impossible for Adams to meet Brown, it ought to be the case that Adams meets Brown. Though the state of affairs in which Adams and Brown meet has become impossible, Adams can still act in ways that respect the value of keeping his promise to meet. For example, he can apologize for not making the meeting, he can try to make it up to Brown, and he can take his appointments more seriously in the future. As this illustrates, people can act in ways that respect the value of states of affairs that intuitively cannot be the case, and thus, they can have reason to act in these ways. So, there is no route, via our assumption about reasons for action, from the claim that something ought to be the case to the claim that it can be case. But there does seem to be a route, via reasons for action, from the claim that an agent ought to act in some particular way to the claim that she can act in that way. So, Harman's suggestion appears to provide one way of explaining why "ought to do" claims have a special connection to connection to claims about what an agent can do.

However, the foregoing explanation depends on the assumption that an agent can have a reason to do something only if she can do that thing, and we might quite reasonably wonder whether or not this is true. As stated, the principle about reasons is subject to an objection. Surely I can have some reason to do something that I cannot do. I have a friend that I would like to visit in California now, so I have *a* reason to go to California. This seems true

even though my budget and my commitments here in New York make a trip to California practically impossible. I would still have this reason to travel to California even if there were a warrant out for my arrest in California, or if I were already imprisoned in New York, or if all means of transportation to California shut down. So, it seems that I can have *a* reason to do something that I cannot do.

But we are entitled to appeal to a more plausible principle about "oughts" and reasons. I am focusing on the implications of *all things considered* normative "ought" claims. The claim that an agent ought, all things considered, to act in a particular way implies more than the claim that the agent has *a* reason to act in that way. It implies that the agent has *most* reason to do that thing. And it is more plausible to suggest that what an agent has most reason to do could only be something that she can do. I have *a* reason to go to California, but if there are obstacles that would make it difficult or impossible for me to go to California, then it seems to be the case that this is not what I have *most* reason to do. If it is truly impossible for me to get to California, then it seems fairly certain that I have more reason to do something else, something that I actually can do. So, we can explain the difference in implications between "ought to do" and "ought to be" by appealing to the principle that what an agent has most reason to do is something that the agent can do.[29]

[29] It should be noted that although I have made use of Harman's suggestion, and although his suggestion is phrased in terms of two senses of "ought," nothing in the explanation above depends on the idea that the sense of "ought" differs in "ought to do" and

So, I have just suggested that the principle, "what an agent has most reason to do is something that that agent can do," provides us with an explanation of why claims that an agent ought to do something, all things considered, imply "can," even while claims that something ought to be the case, all things considered, do not imply that this thing can be the case. This is not meant as a defense of the principle that "ought implies "can." It seems to me that the principle about reasons I am relying on stands or falls with the principle that if an agent ought to do something, all things considered, then that agent can do that thing. At this point, I have only been trying to establish the contours of a plausible "ought"-"can" relationship and, in particular, to show that one who thinks that "ought to do" implies "can" has a good, principled reason for resisting the idea that "ought to be" implies "can."[30]

Finally, there may be another way of supporting the idea that "'ought' implies 'can'" should apply only to claims that say what an agent ought to do. When we use "ought" to express the support of normative considerations, we are presumably evaluating something from among a range of alternatives. Whether we are saying that an action is one that ought to be performed or that a state of affairs ought to be the case, we are saying that it has more

---

"ought to be" claims. Perhaps the sense differs, and this is why "ought to do" claims have a different relationship to claims about what agents have most reason to do. On the other hand, we may suppose that the sense of "ought" is consistent, and that it is what "ought" is being applied to that determines the nature of its relationship to "has most reason to." I do not see any overwhelming reason to favor one view over the other, although I do suppose that there is something to be said for not unnecessarily multiplying senses. For my purpose here, nothing important hinges on which view you take.

[30] Streumer (2007) defends the idea that we cannot have reasons to do that which we cannot do. For discussion of his arguments, see the Postscript at the end of this Chapter.

normative support than its alternatives. Now, a claim that says an agent ought to do something is forward-looking in a certain way. It is a claim about how the agent ought to proceed from the point at which the judgment applies. So, an "ought to do" claim expresses support for a way of proceeding, for a more or less specific course of events from a more or less specific point in time.[31] It expresses support for one of a range of ways of proceeding in time. Claims about what ought to be the case do not seem to be forward-looking in this way. They may be claims about how the world ought to be in the future, but they are often claims about how the world ought to have been, how the world ought to be right now or how the world ought to be at any time. They may even express timeless support for one of a range of various arrangements of the world.

It makes sense that, when evaluating ways in which an agent might proceed, we would count as most favorable only one or more of the ways in which the agent *can* proceed, *i.e.*, only one or more of the ways in which it is open to him to proceed.[32] The claim that an agent ought to do something is an answer to a question about whether it is better for the agent to proceed in

---

[31] Obviously, we want this explanation for past tense claims, claims about what an agent should have done, and there is a very clear sense in which these judgments are not "forward looking." Nonetheless, they express support for what was, at some point in the past, a way of proceeding into the future. In this sense, they are what I am somewhat vaguely calling "forward looking." We can avoid difficulties by thinking of claims of the form "A ought to have done φ" as true if and only if at some point in the past a claim of the form "A ought to do φ" was true. So, these past tense "ought" claims are closely related to present tense "ought" claims, which are less problematically seen as forward looking.

[32] "One or more" because there may be more than one equally good option. It may be a matter of indifference whether one becomes a pediatrician or geriatrician, as long as one does one or the other.

one way or another. It seems reasonable to suppose that it is always better for him to proceed in a way that is open to him, practically speaking, as opposed to one in which he will inevitably fail. But when evaluating ways in which the world might be arranged, I see no similar reason to limit ourselves to the way in which the world currently is arranged or even to the ways it could, practically speaking, be made to be. The claim that things ought to be a certain way is an answer to a question about whether it would be better for the world to be arranged in one way or another, and it seems perfectly sensible to say that it would be better for the world to be arranged in certain way, even if no one could make it so. There may be limitations on the range of alternative arrangements of the world that we could coherently favor. It might be absurd to suggest that a logically impossible arrangement of the world would be better than some possible world. But when comparing and evaluating arrangements of the world we may have good reason to compare not only the ways the world can be, intuitively speaking, but also the ways in which the world might have been, intuitively speaking.

To sum up, I have argued, following Harman, that there is good reason to interpret the principle that "ought" implies "can" as saying only that all things considered, normative claims about what an agent ought to do imply "can." In what follows we will consider whether this limited version of the principle is plausible, and in particular, what sense of "can" could make it plausible.

VI. Irresistible Urges: When We Cannot Help Ourselves.

Section IV presented an objection to the idea that "ought" implies "can,"

and it forced us to get clear about which "oughts" imply "can." In this section, we will consider a different sort of objection, and it will force us to get clear about which "cans" are implied. An example of this second sort of objection can also be found in Sinnott-Armstrong's "'Ought' Conversationally Implies 'Can.'" There, Sinnott-Armstrong mentions, but chooses not to focus on, the following example: "I ought not to laugh at my boss's new haircut, but I can't help myself."[33] It is unfortunate that Sinnott-Armstrong does not develop this example, because it belongs to a family of interesting and instructive cases. To see that, we will have to flesh it out a bit for ourselves.

Suppose, first of all, that Sinnott-Armstrong really means it when he says that he cannot help himself. Uses of the phrase "I can't help myself" often involve some exaggeration. When people break their diets, they often say things like, "I know I should not eat cake, but I could not help myself." Normally, we should probably take this sort of claim with a grain of salt. In the case at hand, though, we should suppose that Sinnott-Armstrong really is making every effort not to laugh, but that the urge to laugh is literally overwhelming. Secondly, let us suppose that, as the phrase suggests, the problem is simply one of controlling oneself. We are not supposed to assume that this person cannot refrain from laughing for the sort of reason that he cannot fly, *viz.* because he lacks the proper anatomy for the task. This person is, in general, physically capable of not laughing. After all, he spends most of the day not laughing, and in many other cases, he does successfully resist the

[33] Sinnott-Armstrong (1984) p.251.

urge to laugh. The problem is that, at the moment he sees his boss, an urge to laugh literally overpowers whatever resistance he can muster up. Given this description, I am inclined to agree with Sinnott-Armstrong. A person should not laugh at someone else's expense, even if he literally cannot help himself.

The situation just described is not that uncommon. It involves a normal person, generally in control of himself, but occasionally overcome by a momentary, irresistible urge. But we should also consider the possibility of agents who have more serious problems with self-control. Consider an example. Winona has the means to pay for the clothes she picks out at Saks, and there is a teller there ready to accept her money. But Winona, let us suppose, is a kleptomaniac. She is overcome by a compulsive desire to take some sweaters without paying. In general, one ought to pay for the clothes one takes from a store, and Winona is no exception. She ought to pay for the sweaters she takes, even though she cannot bring herself to do so. A compulsive disorder drives her to try to steal the sweaters instead.[34]

---

[34] Cases of kleptomania, or other similar compulsions, have been considered in connection with the "ought" implies "can" principle on a number of occasions in the literature, prompting differing reactions in different authors. For example, Peter Vranas has suggested that these cases are not counterexamples because, as a matter of fact, it is not impossible to resist these compulsions. Vranas (2007) pp.183-84. Alternatively, Vranas asserts that if these compulsive disorders truly made it impossible to avoid the compelled behavior then their subject would be "in a certain respect akin to a malfunctioning robot: the concept of obligation does not apply to her." *Id.* at 184. (Query why she is like a *malfunctioning* robot, and not simply like a robot.) In contrast, Peter Graham suggests that he believes that the kleptomaniac provides a counterexample to the principle, but that the availability of Vranas's response can rob the example of its "dialectical punch." Graham (2011) p.340.

I do not pretend to know a thing about the pathology of actual kleptomaniacs. I am skeptical that even an expert could know whether these compulsive desires are resistable or

This example is different than Sinnott-Armstrong's laughing example in a way that might lead some to think that it is not a counterexample. The person who is overcome by an urge to laugh might have little warning that something will prompt this overwhelming urge. He might have been given no warning that his boss has a ridiculous haircut, and he might not have reason to

---

irresistible on particular occasions where they are not in fact resisted. But I do not believe it matters. I see no difficulty in imagining a truly irresistible compulsion and asking what our moral intuitions are about such cases. As indicated above, I have a strong intuition that the kleptomaniac, real or imagined, ought not to steal. Therefore, I find Vranas's first response to miss the point, and regarding his second response, our moral intuitions appear to differ. I attempt to make the case for my intuition above.

Likewise, Graham seeks to support the idea that people with irresistible urges may have an obligation to resist them. Graham argues that we can find support in cases where an action supposedly becomes permissible because of another impermissible action. In one of his examples, I make a promise to you not to tell Melissa where you are. Graham suggests that it becomes permissible for me to break this promise when I learn that you have stolen Melissa's wallet and that she is looking for you in order to retrieve it. Graham argues that the best explanation of this fact is that you have violated an obligation not to steal Melissa's wallet. Finally, Graham contends that this remains true even if we add the assumption that you were compelled, like a kleptomaniac, to steal Melissa's wallet. In other words, I may still break my promise, because, regardless of the fact that you are a kleptomaniac, you were obliged to not steal her wallet. You ought not to have stolen it.

Much of Graham's paper is dedicated to defending the idea that your obligation is the best explanation of why my act becomes permissible. I believe he fails. There is a much better explanation that does not depend on your obligation or on the wrongness of your action. I may break my promise—in some cases, I must break my promise—because (a) Melissa has a right to retrieve her wallet regardless of whether your act was wrong, (b) you have no right to keep it, and (c) Melissa's right trumps my promise. After all, it could become permissible for me to break the same promise if you took Melissa's wallet by accident and you are not even aware that you have it. If the wallet had sufficient value or she had sufficient need, then her right to retrieve her wallet could trump my promise regardless of whether you stole or accidentally acquired it. (Similarly in Graham's organ harvesting example, which I will not explain here, the innocent victims have a right to protection from the organ harvesting surgeon regardless of whether or not the surgeon is a moral agent who is subject to moral reasons and obligations.)

This is not the place to explore Graham's argument any further. The reader can judge for himself or herself whether Graham can defend his claims. It should be noted though that Graham presents this argument in an effort to show that "ought" does not imply "can." Because Graham's examples all rely on the fact that people who cannot help but X may still have an obligation not to X, they provide no counterexample to the version of "'ought' implies 'can'" that I defend in the remainder of this Chapter. Instead, if they could be defended, they would support my version of the principle.

suspect that there would come a time that he wouldn't be able to hold back an inappropriate laugh. So, it is not hard to imagine that there was nothing this person could have done to avoid laughing at his boss's haircut. But people with more serious and systematic motivational disorders, people who suffer from pathological desires and fears, may be quite aware of their problem, and so, they may be able to do what they ought to do by taking special measures to avoid situations in which their compulsion would affect them. So, one might think that what Winona *really* ought to do, all things considered is avoid clothing stores and have someone else do her shopping. In so far as she can do that, she can pay for her purchases and avoid stealing.

It is true that Winona's situation is different than Sinnott-Armstrong's in important ways, but Winona's situation provides a related counterexample nonetheless. First, though it is true that Winona ought to avoid situations in which she will be tempted to steal, it is still true that once she is in those tempting situations, she should not steal. Indeed, she ought to avoid tempting situations precisely because she should not steal. Second, though there are steps that she can take to avoid those situations, and thus avoid stealing, none of the acceptable measures she might take are foolproof. So, we can easily revise the example a bit to show that Winona was not only unable to resist the urge to steal once she was at Saks, but also unable to avoid being at Saks. Suppose that she were brought there against her will, and then set free to do as she pleases. Once set free, she cannot manage to leave without slipping a sweater into her handbag. In this case, she ought to leave the store

without taking anything or she ought to pay for what she takes. Unfortunately, she cannot bring herself to do either of these things.

One might feel some resistance to the idea that the kleptomaniac ought not steal because one supposes that we cannot blame the kleptomaniac for her stealing. The kleptomaniac is not freely choosing to steal; she is the pawn of a desire that is beyond her control. And so, we might reasonably suppose that it is too harsh to hold her accountable for her stealing. But we should be careful to separate our intuitions about what Winona ought to do from our intuitions about whether or not she should be blamed, punished or otherwise held responsible. Responsibility for actions will be discussed in greater depth in the following Chapters. For now, suffice it to say that people are not always blameworthy when they fail to do what they ought. It is quite plausible that a person with a compulsive desire to steal sweaters, who is brought, against her will, to a store filled with nice sweaters, deserves little blame if she steals a sweater. Whether she was dragged there or not, surely her level of blame is mitigated by the fact that her actions stemmed from a pathological psychological disorder. But none of this changes the fact that she should not steal the sweaters. Stealing is morally wrong and often imprudent. Stealing is not "morally wrong *unless you cannot resist the desire to steal*." Moral principles with exceptions of this kind would surely be too lenient.

We can generate countless more examples roughly like Winona's: people with anger management issues should not lash out, but they may not be able to help themselves in certain circumstances; compulsive handwashers

ought to concern themselves with more important matters, but they cannot bring themselves to do so; claustrophobics might find themselves in situations where they really ought to ride the elevator, though they cannot bring themselves to walk through its doors; and smokers and drug addicts ought to kick their habit even if they lack the will power to do so. Of course, these are all examples that involve more extreme problems than Sinnott-Armstrong's laughing case. They involve agents with more serious problems of control.

There is also a range of situations that are less extreme than Winona's, without involving a problem as momentary and fleeting as Sinnott-Armstrong's. Many people find that strong emotional connections make them unable to do what they ought to do. They recognize that they ought to end a relationship or change its normal patterns in some way, and yet they feel that they cannot help but repeat these patterns. While the claims that lovers make are often not literally true, it would be a mistake to think that people affected by love or despair always have full control of their actions. I am not thinking only of young romantics driven to do things that seem silly to an impartial observer, but also of overly faithful spouses and mourning parents. What a spouse who has been wronged, a parent who has lost his child, and an unrequited lover ought to do is highly dependent on the details of the situation, and even when such details are provided, people are sure to disagree about what ought to be done. Nonetheless, it is plausible that, sometimes, people who are heartbroken or mourning get to a point where they ought to take steps toward "moving on," though they are literally unable to bring themselves to do so. We

need not think of these people as suffering from a psychological disorder. They are in the grip of emotions that are beyond their control, though the emotions are ones we can understand and respect.[35]

One might be tempted to conclude that the various examples we have been discussing show that, even when limited to normative, all things considered "ought to do" claims, "ought" does not imply "can." But before we give into that temptation, we should note that these examples are all of a kind. They are all cases in which people have the physical capacity and the opportunity to do what they ought. Their bodies can generally perform the motions that are required. In general, the angry guy's mouth and hands are capable of being at rest, the claustrophobic person can walk through doorways, and the abused spouse is physically able to pack bags and drive off. In addition, nothing in their environment constitutes an insurmountable obstacle between them and what they ought to do: a mad scientist is not forcing the angry man to yell, nothing is blocking the claustrophobe's path to the elevator, and no one is physically restraining the abused spouse. These people are unable to do what they ought to do solely because they are compelled by overwhelming psychological and motivational forces. They are cases in which it is impossible for the agent to do what she ought because she lacks sufficient motivation or control needed for doing what she ought.

---

[35] Graham mentions an example that fits along these lines: "I really ought to put my dog to sleep, but I just can't." Graham (2011) p.357. In many cases, a person saying this would probably be more honest if he said "but I won't" or "but I really don't want to." Nonetheless, I am inclined to agree with Graham that this statement might be truly said by a person with an emotional attachment that makes it truly impossible to euthanize his dog.

Moreover, we treat cases of this sort differently than those in which the agent's inability has some other source. When a person lacks the physical strength, the skill or the know-how for something, we take this as a reason for rejecting the idea that he ought to do that thing. It is false to say that a person who has never learned to swim ought to swim to the rescue of a drowning person. Similarly, when some external obstacle or some feature of the person's situation makes it impossible for a person to do something, this is a sufficient reason for rejecting the claim that he ought to do that thing. If a trained, but off-duty, lifeguard is too far from the water to be of any help to a drowning swimmer or if he is being held at gunpoint, then it is not the case that he ought to save the drowning person.

Given that the cases we have been discussing are all of this one kind, we should not take these cases to show that there is no interesting connection between what a person ought to do and what she can do. Instead, these cases suggest that we need a more carefully limited version of the principle. Inabilities that do not stem from motivational problems do seem to be relevant to the truth of an "ought to do" claim. So, what these cases actually suggest is that if a person ought to do something, then she could do that thing provided she were sufficiently motivated and sufficiently in control. Or, in other words, if a person ought to do something, then she has the physical and mental ability, the skill, and the know-how needed *and* she is in circumstances appropriate

for doing that thing *but* she may or may not have the will to do it.[36]

In the next two sections, I develop and defend this idea. First, I develop a way of categorizing the various factors that make it possible for an agent to do something. This will make it possible to formulate a clear version of the principle that treats motivation and control based inability separately from other kinds. Second, I argue that this principle is the correct version of the principle that "ought" implies "can." It not only avoids counterexamples; it also has the support of certain theoretical considerations.

VII. The Conditions for Action: A Folk View

We will be able to draw the distinctions we need if we first develop a simple theory of how agents are able to act. The view I shall be relying on is deeply rooted in common sense, but it is not normally made very precise. We shall have to develop and clarify each of its elements if we are to have a theory that is rigorous and precise enough for our purposes.

We often pay attention to three different kinds of causal factors that, together, make action possible. Agents have abilities, and if they are in the right circumstances and sufficiently motivated, the exercise of these abilities leads to action. According to this simple sketch of action, agential action is made possible by a confluence of three factors: an agent's abilities, his

---

[36] It has been called to my attention that a similar interpretation of "'ought' implies 'can'" has been suggested by Derek Parfit. *See* Parfit (1984) pp.15 & 506-08. Parfit proposes that, "In the doctrine that *ought* implies *can*, the sense of 'can' is compatible with Psychological Determinism," and that "cannot" entails "ought not" only where "cannot" means "that acting in this way would have been impossible, even if my desires and dispositions had been different." *Id.* at 15. Parfit does not explain why this should be so, although in a footnote he does argue that this interpretation is consistent with certain cases and examples that are taken to support the idea that "ought" implies "can." *Id.* at 506-08.

circumstances or opportunities, and his motives or choices. This common sense conception allows room for luck as well. With some difficult tasks, we think that an agent can have a strong ability, a good opportunity and the needed motivation, but still fail from time to time. A basketball player who is a great free throw shooter will still get unlucky sometimes and miss. In his case, the three factors combine only to make it highly likely that he will make his shot.

The picture of action here is rooted in common sense, but it involves concepts – "ability," "opportunity," and "sufficient motivation" – that require further analysis. In this section, I shall clarify and support this folk view of action by getting clear about each of the crucial concepts in it. There are surely puzzles about some of these concepts that I will not be able to discuss, but I hope to do enough to show that a properly developed version of this view of action can be used to articulate, coherently and helpfully, the causal conditions for action.

Let us start with the notion of an ability. Ascriptions and denials of abilities often serve the same purpose as claims involving "can," "able," and "capable." In some situations, we use "I cannot do that," "I am unable to do that," and "I do not have the ability to do that," interchangeably. But "ability" has one specific usage that suits it for the role it plays in the folk view of action. In this usage, the grammar and structure of ability ascriptions differs from that of claims involving "can" and "able." This reflects the fact that when we attribute (or deny) an ability to an agent we do not merely say that it is

possible for the agent to do something, we do so by ascribing a property to the agent. We say that an agent *has* an ability. So, abilities are properties of agents.[37] Now we need to know more about what sort of properties they are and when we say that agents have them.

Abilities are closely related to causal powers. They are properties that make it possible for agents to do certain things (in the right circumstances). So, when an action comes about through the exercise of an agent's ability, we believe that the agent has made a causal contribution to that event. She has made something happen.[38] This brings out another way in which attributing an ability is different than saying that something can happen or that something is possible. Abilities are abilities *to do* things. So, while it makes sense to say that a person *can*, say, be hit with a water balloon, it would be quite strange to say that a person has the ability to be hit by a water balloon. Similarly, people

---

[37] To be clear, we also attribute abilities to many inanimate and non-sentient entities. A truck might have the ability to haul a load, and a wall might have the ability to withstand a blow. But I restrict my analysis here to the abilities of agents, or more precisely, sentient and purposive agents. The notion of an "agent" in contemporary philosophy seems to me to be largely restricted to that of a sentient and purposive being, and sometimes even further restricted to that of a potentially rational or autonomous agent. In common parlance, "agent" applies more broadly, even to such things as "cleaning agents"—the chemical products, not hygienists or janitors. While I think that this restriction might sometimes warp our vision, it is one that I am following here. That is, I am using "agent" as shorthand for "sentient and purposive agents." We attribute abilities to many things that do not fit the philosophical notion of an "agent," but I shall be focusing on the abilities of agents in this philosophical sense.

[38] I am not claiming here that something that should be described as "agent causation" happens when an agent exercises an ability. A handful of philosophers have special conceptions of "agent causation," and I have no interest in debating the significance of this term here. For a fairly thorough discussion of "agent causation" see O'Connor (2000). Nor am I claiming that when an agent makes something happen by exercising an ability, the buck stops there. Just because the agent makes something happen and is, thereby, intuitively, the cause of that thing, this does not mean that the agent is the ultimate cause of that thing. It does not mean that there is not a causal chain stretching back into time that also causally explains the occurrence of the event.

can get sick, but they do not have abilities to get sick.[39]

Among philosophers who have discussed the notion of an ability, most have suggested that the truth conditions of an ability ascription can be expressed in terms of a conditional.[40] The folk view of action makes this suggestion seem overwhelmingly plausible. According to that view, when an agent has the ability to do something, φ, has a good opportunity to φ, and is sufficiently motivated to φ, the agent will φ, or at least be quite likely to do so. So, it seems to follow that the agent has the ability to φ just in case it is true that if the agent has an opportunity to φ and sufficient motivation to φ, then the agent will (be quite likely to) φ. Of course, this sort of conditional account is only informative if we also have a grip on what it is to have an opportunity and to be sufficiently motivated. Otherwise, it is little more than a restatement of the folk view of action that we are trying to clarify. We could give similar conditional accounts of opportunity and sufficient motivation, but we would be none the wiser for having done so. We still need to know more about what it is to have an ability.

Abilities, I have said, are properties of agents. Opportunities on the other hand are constituted by agent's situations. So, having an ability is a

---

[39] Anselm of Canterbury made a point similar to the one I am making here. He suggests that though Hector can be conquered by Achilles, it would not make sense to say that Hector has the power or the ability to be conquered by Achilles. When ascribing a power or an ability it only makes sense to say that Achilles has the ability to conquer Hector. See Anselm of Canterbury (1998) p.163. Thanks to Scott MacDonald for bringing this passage to my attention.

[40] For some examples of conditional accounts, see Gert and Duggan (1979) p.201 and (1967) p.128, and Vihvelin (2000) p.142.

matter of having a certain causal power, while having an opportunity is a matter of being in certain sort of situation. These facts make it seem natural to think that we have abilities in virtue of being intrinsically constituted in a certain way, and that we have opportunities in virtue of our relations to things around us. While the simplicity of this suggestion is tempting, it is not borne out in our actual attributions of abilities. Some of our abilities are grounded in our (seemingly) intrinsic properties. I have the ability to wiggle my finger just because of how I am anatomically constituted. I might lack the opportunity to do so because my hand is momentarily caught under the wheel of a car, but as long as the relevant parts of my body remain as they are intrinsically, I should think that I would maintain the ability to wiggle my finger. But we also have abilities that we could only possess in virtue of certain kinds of extrinsic properties. I have the ability to buy clothes, for example, only because there is money in my bank account or credit on my credit card and a generally accepted system of accepting money or credit as payment for goods. A sitting president has the ability to veto a bill only because he occupies the office, and the U.S. Constitution grants that power to the president. I lack the ability to write a note to my wife because I am unmarried. So, intuitively plausible ability attributions suggest that abilities can depend not only on our intrinsic constitution, but also on some of our extrinsic properties. This fact complicates matters. We cannot distinguish abilities and opportunities in terms of the intrinsic/extrinsic distinction. However, with a better understanding of how extrinsic factors can ground abilities and how they

contribute to opportunities, we can properly distinguish ability from opportunity.

I have stressed that the abilities I am concerned with are abilities to do things. But when we do something, it is never the case that there is only one way of describing what it is that we do.[41] Suppose that Obama, while grasping a pen, made a certain motion with his hand. Suppose that in doing so, he signed his name. And in doing so, he made a bill into a law. And in doing so, he authorized the overhaul of financial regulation. What the President does depends not simply on the changes in his intrinsic state, but also on his extrinsic properties. He may make the same motion with a pen in many circumstances. Sometimes when he does this, he will be making a law, and other times he will be signing a check. If he were to make the motion on the last day of his term, he might be signing a bill into law, but if he were to wait until the following day, he certainly would not be doing this.

Now, let us focus on the most basic description of the action that Obama performs. What I mean by this is that we may describe what Obama does simply in terms of the changes to his intrinsic state. This will involve simply the movements of his bodily parts, and perhaps, some of his mental activities.[42] Let us call Obama's action, so described, "moving his hand just

---

[41] Some disagree about whether there is one action here, described in different ways, or many distinct actions. I follow Davidson in speaking of this as one action that has many different descriptions, but nothing I say should depend on this. The proponent of the multiple action view should be able to accept the point that I am going to make, given the necessary changes in wording. For discussion of this issue, see Davidson (1971) and (1967).

[42] If externalism about mental content is right, then many simple mental activities will depend not only on one's intrinsic properties, but also on one's relations to the environment. The ability to think about water, for example, will depend on one's having been properly

so." It seems to me that Obama has the ability to move his hand just so in virtue of his intrinsic properties alone. His muscles, ligaments, nerves, and such are configured in a certain way, and, through training, his brain and body has become configured so that, when he chooses to move his hand in this way and when there are no abnormal obstacles, he does move his hand just so. Abilities to perform actions that are basic in this sense are abilities that agents have just in virtue of their intrinsic nature. Let us call them "basic abilities."

Of course, describing actions in these basic terms is a good way to miss out on much of the significance of what we do. Much of what we do simply is not captured by these basic descriptions of the motions of our bodily parts, but instead must be captured in non-basic descriptions of our actions. Which non-basic descriptions apply to our actions depends not only on the things that we do with our body, but also on the social and environmental setting of our body. And so, I shall argue, our abilities to perform actions of these kinds are dependent on or grounded in our relations to our physical and social environment.

Before I can make this fully clear, there is one other feature of abilities that we need to note. Ability ascriptions typically involve what are sometimes described as "incomplete predicates."[43] We say that people have abilities such as the ability to run a 4-minute mile, or to jump 6 feet, or to type 80 words

---

related to a watery environment.

[43] See, for example, Prior (1985) pp.6-8.

per minute. But the explicit wording of our ability ascriptions does not usually say precisely the kind of action that we are saying that the agent is able to perform. When we say that someone has the ability to run a 4-minute mile, we do not mean that they can do so in any circumstances whatsoever. Even the select few who have the ability to run a 4-minute mile can do so only in a very limited range of circumstances. The fact that they cannot run a 4-minute mile while wearing hiking boots or on uneven ground or just after eating dinner does not show that they lack the ability to run a 4-minute mile.

This might tempt one to say that when we ascribe the ability to run a 4-minute mile we mean that the person has the ability to run a 4-minute mile *in favorable circumstances*. But this is not very helpful. "Favorable circumstances" could mean nearly anything. I can run a 4-minute mile in circumstances where I have rocket-propelled shoes or in circumstances where the muscles in my legs have been altered. The moon would provide me with a favorable circumstance for jumping high enough to dunk a basketball, but the fact that I can dunk in such favorable circumstances does not show that I have the ability to dunk a basketball.

When we say that someone has the ability to do something, like run a 4-minute mile or dunk a basketball, we mean something more specific than what we say explicitly. In most cases, the full content of the ability we ascribe to a person is given implicitly by the context and by our shared understanding of the normal conditions for the activity explicitly mentioned. So, when I say that someone has the ability to run a 4-minute mile, what I mean is something

more like the following: he has the ability to run a 4-minute mile, *given adequate rest and health, on a track surface that meets the modern standard, while wearing suitable clothing and footwear, while not held back by any physical restraints or weights, while being affected by the Earth's gravity, and while breathing clean air with oxygen concentration roughly similar to that at sea level.* This is closer to an accurate and explicit statement of the property that I mean to be attributing to the agent, but it is no doubt still incomplete. It is not clear that we could ever state explicitly the full and specific content of the ability that we mean to ascribe or deny to an agent. Fortunately, we do not need to, since, most of the time, we share a common understanding of many features of the content.

Sometimes we do go out of our way to make some part of the content of an ability explicit. We do so when we mean to ascribe an ability that differs from what we would otherwise expect people to understand. So, someone might have the ability to run a 4-minute mile on a slight downhill, but not on a flat. Or, someone might have the ability to type 80 words per minute on a full-size keyboard, but not on a smaller laptop keyboard. In these cases, much of the content of the ability is still given implicitly by our shared understanding, but it is modified by the explicit qualifications and changes to that normal expectation. I shall say that the implicit and explicit qualifications on the content of an ability provide us with the *fully specified content* of an ability.

The two points just covered—the fact that actions admit of different descriptions and the fact that the full content of ability predicates typically

involves a great many implicit qualifications—allow us to draw a distinction between what it is to have an ability and what it is to have an opportunity and to see when abilities are grounded in an agent's extrinsic properties. Take any standard ability predicate, like "the ability to φ." The fully specified version of this predicate is given by "the ability to φ in circumstances that meet conditions, C," where C is the long list of conditions and qualifications that specify the sort of situation which we normally understand to be appropriate for φ-ing. Having an opportunity to φ, or having an opportunity to exercise one's ability to φ, is simply a matter of being in a situation where those conditions, C, are met.

It may help to clarify if we take a look at a specific example. Consider, again, the ability to run a 4-minute mile. I suggested an approximation for the fully specified content of this ability. Given that specification, what I am suggesting now is that a person has an *opportunity* to run a four minute mile when she is adequately healthy and rested, on a track surface that meets the modern standard, wearing suitable clothing and footwear, not held back by any physical restraints, affected by the Earth's gravity and breathing clean air with normal oxygen content. The considerations that fill out the content of a fully specified ability are the factors that determine whether or not someone has the opportunity to perform the relevant action. So, what counts as an opportunity to do something is given by the, typically implicit, conditions that make up the fully specified content of an ability.

The multiplicity of action descriptions is important as well. Sometimes,

when a person runs a 4-minute mile, it is also true that he wins a race. Whether or not running a 4-minute mile is also winning a race depends on a number of factors that are extrinsic to the agent. It depends most obviously on whether the agent is running in a race and on how fast the other racers run. Now, we may want to know whether or not the agent has the ability to win a particular race. I believe that we will say that he has this ability when he has the ability to do something else, like run a 4-minute mile (in the circumstances C), *and* he is extrinsically such that his doing this other thing amounts to, or is reasonably likely to amount to, winning the race. His ability to win the race depends on his ability to run at a certain speed over a certain distance as well as the fact that there is a race going on and the other runners are not very likely to run any faster than him.

Let us generalize this suggestion. A person has certain basic abilities (*e.g.*, abilities to engage in certain bodily movements) in virtue of his intrinsic properties. He has an opportunity to exercise one of these basic abilities when his circumstances meet the conditions given by the fully specified content of that ability (*e.g.*, when external restraints do not prevent the relevant bodily motions). When a person exercises one of these basic abilities and performs a basic action, he may also do something else that is not basic in this way. So, for example, repetitively placing one foot in front of the other at a pace sufficient for covering a mile in four minutes and doing so for four minutes may also be, in the right circumstances, winning a race. Whether or not performing the basic action also amounts to performing a non-basic action

will depend on the circumstances that the agent is in. So, an agent has an ability to perform a non-basic action, NB, when she is intrinsically such that she has the ability to perform some basic action, B, and she is extrinsically such that doing B is, or is reasonably likely to be, doing NB.

Given this, we may propose the following accounts of abilities:

> An agent, *A*, has a **basic ability** to perform a basic action, φ, if and only if *A* is intrinsically such that if *A* were sufficiently motivated to φ and *A* had an opportunity to φ, then *A* would be relatively likely to φ (taking into account the difficulty of φ-ing).
>
> An agent, *A*, has a **non-basic ability** to perform an action, φ, if and only if:
> 1) *A* has certain basic abilities to perform certain basic actions $\varphi_1, \ldots, \varphi_n$; and
> 2) *A* is extrinsically such that by performing an appropriate subset of these basic actions, $\varphi_1, \ldots, \varphi_n$, in an opportunity to φ, *A* is relatively likely to h; and
> 3) if *A* were sufficiently motivated to φ, *A* would attempt to do so by exercising some appropriate subset of her abilities to $\varphi_1, \ldots, \varphi_n$.

I will clear up a few of the details of this account in a moment. First, let me offer the following account of an opportunity.

> An agent, *A*, has an **opportunity** to perform an action, φ, if and only if *A*'s circumstances are such that they meet the conditions given in a full specification of the ability to φ.

This account of opportunities applies whether we are talking about a basic or a non-basic ability.

The account of abilities just given relies on a notion of being "sufficiently motivated." Thus, the account depends on cashing this notion out in the right way. I say that an agent is sufficiently motivated to φ when her motivation to φ

is sufficient to determine her will and result in her trying to φ in spite of any motivation she has to not-φ and, more generally, any psychological resistance she has to φ-ing. So, an agent will be sufficiently motivated to φ when she wants to φ more than anything else (or when she wants to φ at least as much anything else and she makes an arbitrary choice to φ) and she does not suffer from any weakness of will. This account of sufficient motivation does not depend on setting some "amount" of motivation as the minimum threshold. Indeed, it does not depend on the possibility of quantifying motivational forces. A person who whimsically walks onto an elevator, but could have been quite content to wait for the next one, seems to be subject to a weaker motivational force than the claustrophobic person who is struggling unsuccessfully to overcome her fear. The claustrophobe may be setting all the force of her will against her fear, but she is still insufficiently motivated. Meanwhile, it may be that the whimsical person would have lost interest in riding the elevator given the slightest distraction, yet because this does not happen, he counts as sufficiently motivated.[44]

The account above also makes use of probability language. It claims that there is a probabilistic connection between ability, opportunity and

[44] Because my account relies on a counterfactual which hinges on the agent's motivation, one might suppose that my account of abilities fails in Frankfurt scenarios or their analogues, *i.e.*, in cases in which an intervener stands ready to prevent the successful exercise of a certain ability, say the ability to run, whenever the agent becomes sufficiently motivated to run. This would be a worrisome objection only if we would consider such cases to be ones in which the agent has an opportunity to exercise the ability in question. But it is clear that Jones does not have an opportunity to run if Black stands ready to shoot or to tackle Jones anytime he becomes motivated to do so. I see no reason to think that the case would be any different if Black produced the same result by installing some sort of device in Jones's brain, as in the typical Frankfurt scenario.

motivation, on the one hand, and success, on the other. It would be a clear mistake to tie ability and success too closely. People have all sorts of abilities that do not guarantee them success even in the best of conditions. The clearest examples come from sports. Take free throw shooting in basketball. When he was an active player, the Knicks' Allan Houston was one of the best free throw shooters in professional basketball. He consistently made over 90% of his free throw attempts, which is to say he consistently missed almost 10%. It is not plausible to say that these misses can be accounted for by a lack of opportunity or a failure to try. The conditions of his free throw shooting are extraordinarily consistent. His shots took place in the context of a sanctioned National Basketball Association game, with a regulation ball and at a regulated distance from a regulation hoop. His missed shots were also more or less regularly interspersed among his attempts. So, it is unlikely that they were due to a temporary lack of health or a lack of interest or motivation. Instead, it is much more plausible to say that he has a very strong ability to make free throws, but that given the difficulty of making free throws, people with the ability will miss from time to time.

This last example is consistent with the idea that ability, opportunity and motivation do not guarantee success, but they make it highly probable. However, even this would be too restrictive. Consider baseball players for a moment. The very best home run hitters in baseball actually hit home runs in very few of their chances at bat. Alex Rodriguez, for example, has been one of the best home run hitters in major league baseball. He was on pace to

surpass the career home run record. In short, there can be no denying that he has the ability to hit home runs. Yet, he turns only a very small minority of the pitches he sees into home runs. To be fair, not every pitch is hittable. But even if we restrict our attention to pitches that provide him with a reasonable opportunity for hitting a home run, say, pitches in the strike zone, he still hits only a very small minority of them for home runs. Hitting a home run in professional baseball is a difficult thing to do, and so, even people with the ability to hit home runs will succeed fairly rarely.

These two examples suggest that the rate of success (given opportunity and motivation) associated with having an ability depends upon the difficulty of the task at hand. For some very easy tasks, it may be roughly one hundred percent, but for very difficult ones, it may be far below fifty percent. There is no general formula that will tell us how often a person with the ability to φ should succeed. We can simply say that it should vary with the difficulty, and that at a minimum, having an ability to φ must make a positive contribution to the likelihood of φ-ing. In other words, a person with the ability to φ should be more likely to succeed at φ-ing than someone lacking the ability, someone who is relying purely on luck. So, even if one has a weighted coin that has a 90% chance of landing tails, one does not have the ability to make it come up tails as opposed to heads. For one cannot make any positive contribution to its odds of coming up tails rather than heads.

The examples just discussed suggest that we should not expect too close of a connection between the exercise of an ability in the right

circumstances and success. However, the mere fact that people exercising abilities exhibit success less than one hundred percent of the time does not show that the connection between ability and success is accurately or helpfully captured in terms of probability. The presence (or absence) of an ability is often equally apparent in a person's successes and his failures. When Alex Rodriguez hits a foul ball or misses a pitch altogether, he still displays evidence of his ability to hit home runs. He approaches batting in a certain way, carefully discriminates between good and bad pitches, and swings with form and timing that are, in general, apt to lead to home run hitting. A competitor who has gotten the better of his opponent in each of their meetings may still be aware that his opponent has the ability to best him; he can know this because even in failure, his opponent displays some properties that serve as indicators of his ability. An account that simply says exercising an ability in the right circumstances makes success relatively probable might seem to offer very little by way of an explanation of these facts. One might suggest that the mark of an ability to φ is not the fact that a person is reasonably likely to φ when he tries in the right circumstances, but that he displays some state that is appropriately linked with successfully φ-ing.

There are two points to make in defense of the probabilistic account that I have suggested. First, it is not clear how to cash out this talk of "displaying a state that is appropriately linked with successfully φ-ing" except in probabilistic terms such as, "displaying a state that is likely (relative to some context sensitive standard) to result in successfully φ-ing." If this is an

accurate way to understand the suggestion, then there will be very little difference between my account and an alternative along these lines. It will remain true that when a person with the ability to φ, tries to φ in the right circumstances, she is relatively likely to φ. Second, my account does an adequate job of explaining the fact that there can be clear evidence of an ability even in failed attempts to exercise the ability. For basic abilities, my account makes explicit that the conditional—if an agent with the ability to φ is sufficiently motivated to φ and has an opportunity to φ, then she is relatively likely to φ—holds in virtue of how the agent is intrinsically. Thus, it suggests that the agent is in some intrinsic state that makes it the case that he is reasonably likely to φ, whether or not he does succeed in φ-ing. When one exercises a non-basic ability to ψ, one does so by exercising a basic ability to φ that is likely but not guaranteed to result in ψ-ing. So, for example, when Alex Rodriguez attempts to hit a home run, he does so by exercising more basic abilities such as the ability to swing the bat in a certain way that is apt to produce home runs. When Allan Houston missed a free throw, he still typically coordinated his movements in a way that constitutes good form for shooting free throws. According to my account, he has the ability to make free throws in virtue of having this more basic ability to coordinate a number of bodily movements in a certain way, and he can successfully exercise this more basic ability whether or not he thereby makes his free throw. So, at worst, my probabilistic account can track the truth of ability ascriptions just as well as the alternative suggestion, but it seems that, better than that, my account does a

good job of explaining the evidence of abilities in both successes and failures.

This, I believe, is enough to conclude that the folk view of action can be developed into a coherent and principled way of articulating the causal conditions of action. The view that I have developed distinguishes three distinct kinds of factors needed to produce action: ability, opportunity, and motivation. In the next section, I use this account to give a precise version of the principle that "ought" implies "can," a version that is supported by the intuitions explored by the previous section, and I develop arguments to support this version of the principle.

VIII. Why "Ought" Implies Ability and Opportunity, Not Motivation:

In Section VI, we saw that when a person is unable to do something because she lacks the ability or the opportunity to do that thing, we are inclined to reject the idea that she ought to do that thing, but when a person is unable to do something because she is unable to overcome motivational obstacles to her doing the right thing, *i.e.*, when she is not sufficiently motivated to do that thing, we do not take this to undermine the fact that she ought to do that thing. In Section VII, we drew distinctions between abilities, opportunities, and motivation. These distinctions allow us to offer a principle that reflects these intuitions:

> If, all things considered, a person ought to do something, φ, then that person has the ability to φ and an opportunity to φ.

Now we need to consider what can be said in favor of this principle.

In addition to the support of the cases we have discussed, the principle

above has the support of the following conceptual argument. The ethical thing for a person to do in a particular situation *just is* the thing that she would do in that situation were she sufficiently motivated by the relevant ethical considerations. Similarly, the rational thing for a person to do in a particular situation *just is* the thing that she would do in that situation, if her will was determined by rational considerations. The equivalences here can be generalized: all things considered, the proper thing for a person to do in a particular situation just is the thing she would do in that situation, if she were sufficiently motivated to act in accordance with the balance of all applicable normative considerations. So, what a particular person ought to do, all things considered, in a particular situation is the thing that she would do if she were motivated in accordance with the balance of all applicable reasons.

What would a particular person do in a particular situation if she were sufficiently motivated in accordance with the balance of all applicable reasons? To answer this question, we should imagine a situation that is as close to the actual scenario as possible, while making only the adjustments required by the antecedent of the conditional. In other words, to answer this question, we should hold fixed this person's abilities and opportunities, but imagine that she is motivated in accordance with the balance of reasons that apply to her and then consider what she would do. Anything that she would do would have to be something that she actually has the ability and opportunity to do, but it may be something that, in the actual scenario, she is not sufficiently motivated to do. So, for any way of acting, φ, if our person

ought to φ, then she must have the ability and opportunity to φ.

This argument is, to my mind, compelling. It depends on a plausible conceptual connection between what a person ought to do and what that person would do if she were properly motivated. The existence of this connection is clearly demonstrated by the useful role that the idea of an agent who is idealized in just this way can play in our practical reasoning. People who hold a wide variety of moral theories, or who hold no particular moral theory at all, frequently rely on character models to aid their thinking about what they ought to do. Even those who do not choose to rely on this sort of device should admit that, provided we pick the appropriate model, it is a valid way of answering the question, "What should I do?" They should also agree that the proper role model is one who is not significantly different in terms of abilities and circumstances, but who has the right sort of character and, thus, is properly motivated.

The principle that I am recommending reflects a reasonable conception of the role and purpose of claims about what one ought to do. These claims are or should be practical and action guiding while also promoting value. A theory that suggested that an agent ought to do things that go beyond his abilities or that are impossible in his circumstances would be an impractical and unreasonable theory. It would underwrite impossible demands and offer little in the way of guidance. At the opposite end of the spectrum, a normative theory that suggested that a person ought to do something only if she can actually work up the motivation to do it would be nearly pointless. It would not

demand that we act virtuously. Instead, it would make the demands of morality and rationality depend on what the agent already cared about and on how stuck the agent is in her ways. This would offer us little in the way of guidance because it would demand too little. Instead of either of these alternatives, the demands of a moral theory should be reasonable, prescribing only actions that agents have the ability and opportunity to perform, but they should also be demanding, expecting us to be better people than we often are.

It might be thought that "ought" claims are stronger than this: that they are not merely practical and action guiding, but that they place some stronger form of obligation or demand upon an agent, and as such it would be unfair to impose them without being more sensitive to how an agent could possibly be motivated. But it is not clear what this stronger sense of "ought" is supposed to be. Linking this sense with obligations does not help to make it clear, for the practical, action guiding "ought to do" claims I have in mind very often express obligations. Winona, for example, has an obligation not to steal, and so she ought not to do so. The only sense in which "ought" could express some stronger normative claim, one for which it might be unfair to fail to consider motivational obstacles, is one that is tied to blameworthiness, such that "A ought to φ" more or less entails that A would be blameworthy if he failed to φ. This is not a use of "ought" that I am familiar with or that I find particularly useful. Recall that I am inclined to say that Winona ought not steal, but that she may not be blameworthy if she does. Separating what an agent ought to do from what she would be blameworthy for doing reflects the

fact that there are excuses, *i.e.*, considerations that may mitigate or entirely remove blame without implying that the act in question can be justified. And if we need a term that tracks whether or not an agent would be blameworthy for φ-ing, I prefer "would be blameworthy for φ-ing." Of course, if one understands and sees a use for stronger "ought" claims of this sort, they may use them. I am concerned with "ought" in its action-guiding use, which I believe includes its obligation-imposing use. What I have argued is that this action-guiding usage should guide our will, not be guided by the will that we already have.

There is one final complication that must be addressed. I have been arguing that some fairly basic considerations about the nature and point of a normative theory suggest that, in determining what an agent ought to do, it is appropriate to set motivational considerations aside. But I must also admit that it would be too much to set all motivational considerations aside. After all, I noted earlier that normative reasons only apply to an agent that has some capacity to appreciate the force of those reasons. A cat does not have a moral duty to respect the rights of human beings, because a cat is not capable of appreciating and guiding her actions in accordance with moral considerations. At least part of what the cat is lacking is a capacity to be motivated to act on moral reasons. And it seems that just in virtue of this it is always false that a cat ought to do something for moral reasons. So, we might wonder why, if we are supposed to set aside the motivational limitations of, say, a kleptomaniac and say that the kleptomaniac should not steal, we should

not set aside the limitations of a cat and say that the cat ought to respect the rights of persons by not scratching them.

The first thing that we need to do is get clear about this proposed restriction on what an agent ought to do. It is intuitive to suppose that for an agent to have reason to do something, φ, she must at least be capable of coming to appreciate the reason. She must, at least, have the potential to appreciate the relevant reasons. This does not mean that it need be easy for anyone to make her appreciate and care about those reasons. Depravity may make it extremely difficult to get a person to see moral reasons as legitimate reasons, but this is not enough to show that the depraved person should not act morally. If the depraved person retains the capacities that would make it possible for her to care about moral considerations (or, for example, to see people, among other things, as deserving a certain kind of respect), that is all that is required.

In order to have a reason to φ, an agent must have the potential for appreciating that reason. Of course, having the potential for appreciating a reason must not amount to actually appreciating the full force of that reason and being motivated in accordance with that reason's weight. Requiring this would run afoul of our intuitions that the kleptomaniac ought not steal. Our examples of people who, because of motivational obstacles, cannot do what they ought to do were all meant to be examples of people who could appreciate the considerations that supported what they ought to do. But, while they have the capacity to appreciate these kinds of considerations and while

they may have some motivation to act on them, they do not because they are driven by some stronger motivational force.

What we should require is the following. In order for an agent to have a reason to φ, and so, in order for it to be the case that this agent ought to φ, it must be the case that this agent could, in principle, come to care about that sort of reason, though this might require extensive education, training, conditioning, intellectual persuasion, and exposure to the value underlying that reason. This may only be possible "in principle" because, for example, in particular cases there may be no one around who has the resources, the skills, or the motivation to provide the kinds of education that would be necessary. All that is required is that the agent has the basic intellectual and emotional faculties such that the right sort of education would be effective in producing appreciation for the relevant reasons. When this is the case, an agent has a sufficient potential or capacity to appreciate a reason.

Applying this suggestion, we might say the following. People of reasonable intelligence can in general come to appreciate and care about a wide range of reasons: moral, prudential, etc. They can come to appreciate the value not only of themselves and others who are similar to them, but also of people very different than themselves, of animals, of natural areas, of aesthetic beauty, and many other kinds of things. Intelligent people who suffer from weakness of will, from compulsive desires, from pathological fears, and so forth, can still come to appreciate and care about this full range of reasons. They simply have difficulty controlling their actions in accordance with these

considerations. On the other hand, there are other people who lack the potential to appreciate certain kinds of value and certain kinds of reasons. Insanity, senility, or a mental defect may make a person incapable of appreciating a wide variety of reasons for action. There might be no amount of moral education that could prompt a mentally incompetent person to appreciate moral value. Similarly, non-human animals may only be capable of appreciating a limited range of reasons. Some animals can be made, through training, to see that violence towards humans is not in their own interests. But it seems doubtful that even the most intelligent animals can come to appreciate what we might call "moral reasons" for not acting violently. So, I suppose that if it is true that a dog ought not bite, it is not for moral reasons.

So, it seems that if an agent ought to do something, then she must not only have the ability and opportunity to do that thing, but also the potential for appreciating the reasons why she ought to do that thing. She must also have some capacity that would make it possible for her, given the right sort of education, to feel the force of the relevant reasons. This is problematic in so far as it prompts one to wonder why, if the truth of "A ought to φ" requires that A has this potential to feel the force of the relevant reasons for φ-ing, it should not also require that A is actually motivated by them. The considerations I have been relying on, concerning the role of a normative theory and the importance of the concept of an agent who is properly motivated, are sufficient to answer this question. A normative theory must respect the equivalence between what an agent ought to do and what a properly motivated version of

that agent would do. Of course, the properly motivated version of that agent must still be a version of that agent. It should be like that agent except that its will is in proper order. But to imagine a version of the agent that appreciates reasons that the actual agent does not even have the potential to come to appreciate is not to imagine the agent with its will in proper order. It is to imagine the agent with an entirely different kind of will, a will that is susceptible to considerations that the actual agent could not possibly see as reasons at all. A normative theory should demand that the agent be properly motivated, relative to that agent's potential.

The considerations presented in this section suggest that if a person ought to do something then she has the ability and the opportunity to do that thing, and that she has the potential to appreciate the reasons that support doing that thing. In general, when we focus on normal, healthy human agents, the last condition can be ignored. We can safely assume that normal, healthy human agents can be taught to appreciate the full range of considerations that we are aware of: moral, prudential, grammatical, etc. But normal, healthy humans vary a great deal in their abilities and their circumstances. In so far as they do differ, the things that they ought to do will also differ. What an agent ought to do is what she would do if she were an ideally motivated human agent who had to cope with the actual agent's limited abilities and opportunities.

IX. The Argument from Determinism Revisited:

Let us return to the Argument from Determinism and evaluate the worry

that if “ought” implied “can,” then determinism would imply that none of us ever do something we should not do. The argument that supported that conclusion ran as follows:

1) If determinism is true (and people cannot change the laws of nature, and people cannot change the past in any significant way), then no person can ever do anything other than what he or she actually does.

2) If a person ought to do something, then he or she can do that thing.

3) So, if determinism is true, then it is never the case that a person ought to do something other than what he or she, in fact, does.

In discussing this argument and in discussing the meaning of “can,” we saw that we must interpret Premise 1 as saying that if determinism is true then no person can do anything other than what he actually does *in the sense that* doing anything else would be incompatible with the laws of nature and a full description of the state of the world at any point prior to that person’s action. Alternatively, according to Premise 1, if determinism were true, then doing otherwise is incompatible with the laws of nature and a full specification of the range of causally relevant features, even prior to the agent’s making a decision.

There are two questions to ask now. Have I argued in this paper that Premise 2 is true, that if a person ought to do something, then she can do that thing? And if so, is it true in a sense that makes this argument valid? If we take Lewis to be right about the meaning and use of “can,” then I think we

should say that I have defended Premise 2 in this paper. Recall that Lewis suggested that we use "can" to say that something is compatible with some range of relevant facts. I have here argued that if an agent ought to do something, then doing that thing is compatible with the agent's abilities and opportunities. There is no reason to suppose that in saying that an agent can do something, we could not use "can" with a focus only on considerations about the agent's abilities and opportunities. What I have argued is that if an agent ought to do something, then she can do that thing in the sense that doing that thing is compatible with her abilities and her opportunities. Of course, I have also argued that doing that thing need not be compatible with considerations about the agent's motives. So, while it is true that "ought" implies "can," it is not true in the same sense that determinism implies "cannot." Determinism implies that doing otherwise is incompatible with the full range of causally relevant features of a situation. "Ought" implies only that performing the relevant actions is compatible with ability and opportunity, not with one other crucial, causally relevant feature, motivation. So, "can" is used with different significance in Premises 1 and 2, making the argument invalid.

It may be helpful to provide one quick example that shows how the argument fails. Suppose that I find a misplaced wallet containing a significant amount of cash and a valid drivers license, and suppose I pocket the cash before turning the wallet over to the police. If determinism were true, then there is a clear sense in which I could have done nothing else. Doing anything else would have been incompatible with the laws of nature and the state of the

world at anytime in the past. But it remains the case that I ought to have left the cash in the wallet and made sure that the wallet made it back to its rightful owner. Doing that was certainly compatible with my abilities. I had the ability to simply close the wallet and walk it to the police directly, and I had a clear opportunity to do so. I took the money because I was motivated by greed. If I had been motivated differently, if I had been sufficiently motivated to do the right thing, then I certainly would have left the cash in the wallet. I could have left the cash in the wallet in the sense that it was compatible with my abilities and opportunities (though not compatible with my will). So, in spite of being deterministically caused to take the money, there does not seem to be a reason to deny that what I should have done, what would have been the right thing to do, was to leave the money.

So, we may maintain that "ought" implies "can" without being driven to think that determinism would undermine the truth of a whole range of plausible "ought" claims. And this, of course, is one last thing to be said in favor of the idea that "ought" implies "ability and opportunity." It is not plausible to suppose that if the world happens to be deterministic, then it is never the case that we ought to do something other than what we in fact do. Surely, I should not have drunk so much the other night. Surely, I should not have been so critical of my roommate yesterday. I am certain of both of these facts. But certain as they are, they do not seem to me to be the right kind of facts to support any position on the truth of determinism. To assume that determinism is false on these grounds would be to make a metaphysical assumption, without any

metaphysical basis, simply to uphold our moral intuitions. Instead, we should prefer an explanation of these normative claims that does not make them dependant on indeterminism, and that is what my account provides.

X. Postscript

In the years since this Chapter was written a number of articles have been written on the principle that "ought" implies "can" and related topics, one of which merits some attention here. In "Reasons and Impossibility," Bart Streumer argues that a person cannot have a reason to do something that he cannot do, and that this is the reason why "ought" implies "can." His central thesis is similar to the claim I make in Section V above to explain why the "ought" implies "can" principle should be limited to "ought to do" claims. There, I suggested that all things considered "ought to do" claims entail claims about what a person has most reason to do, and that a person cannot have most reason to do something that he cannot do. Streumer makes a stronger claim, that a person cannot have *any* reason to do something that he cannot do. This is a claim that I reject in Section V. However, Streumer also acknowledges that he may have to abandon this stronger claim and concede that it is sufficient for his purposes if a person cannot have "a reason to perform an action that is not outweighed by other reasons if it is impossible that this person will perform this action."[45] I take this idea to be roughly equivalent to my claim in Section V, and thus, if correct, his arguments should support the claim I make there.

[45] Streumer (2007) p.359.

I do not want to evaluate Streumer's arguments in detail. There are reasons to quibble with each, and they do not necessarily entail his conclusion. But they each have some persuasive force, and I have little to add to his defense of them. I would be inclined to simply point the reader to them for further support of my claims here. However, Streumer does not recognize that the sense of "can" in "ought" implies "can" is one that refers to having an ability and an opportunity, not necessarily the motivation. Thus, Streumer believes that his arguments lead to the conclusion that a person cannot have reason to do something he cannot be motivated to do. Indeed, Streumer even accepts that if a psychotic murderer cannot resist the urge to murder, there cannot be a reason for him to stop murdering.[46] I find this to be highly implausible, as my argument in this Chapter should indicate. Therefore, I simply want to explain that even if his arguments succeed, they do not require this implausible conclusion. They are perfectly compatible with my version of "ought" implies "can."

Streumer's first argument, which he calls the "Argument from Crazy Reasons," is that unless a person's reasons are limited by what he or she can do, then people would have reasons to do all sorts of bizarre and impossible things, like travel back in time to single handedly stop the Crusades, or jump 30,000 feet in the air to single handedly repair a jet's failing engines, or invent cures for diseases even though one cannot grasp the science required to do so. Tellingly, all of Streumer's examples of implausible reasons are reasons to

[46] *Id.* p.370.

perform actions that are not within our abilities and opportunities, not actions that we simply lack the motivation to perform. There is nothing implausible about having a reason to do something that we are unmotivated to do. It may be "crazy" to say that a person has reason to jump 30,000 feet up in the air, but there is nothing "crazy" about suggesting that a kleptomaniac has reason not to steal or that a person standing in the road paralyzed with fear has reason to move.

Streumer's second argument, the "Argument from Tables and Chairs," is that when a person cannot perform a particular action, that person is "in the same position with regard to this action that a table or chair is in with regard to all actions." Thus, Streumer argues that, like the table or chair, the person has no reason to perform that action. Of course, tables and chairs lack abilities to perform any actions. Moreover, they are not simply unmotivated to perform actions, they are entirely incapable of appreciating any normative considerations or becoming motivated by them. Thus, while a person who lacks the ability or opportunity to perform a particular action might be said to be in some manner akin to a table or chair with regard to that action, it does not follow that a person who does have the ability and opportunity to perform the action in question and who has the general capacity to appreciate reasons is in any manner akin to a table or chair. Unlike a table or chair, this latter sort of person would perform the action (or would try and be reasonably likely to succeed) if he appreciated and was sufficiently motivated by a reason to do so. Therefore, this argument does not suggest that a person cannot have a

reason to do something if he is pathologically motivated to do otherwise.

Finally, Streumer's third argument, the "Argument from Deliberation," is that if we could have reason to do something impossible, then rational deliberation would be pointless because it would almost always lead to the conclusion that we have most reason to travel back in time to prevent serious atrocities that involved huge amounts of death and suffering, such as the Crusades, slavery, and the two World Wars. Deciding what we have most reason to do would be pointless if we were constantly led to the conclusion that we have most reason to do things that we clearly lack the ability or opportunity to do. Is the same true if deliberation sometimes leads us to the conclusion that we should do things that we cannot successfully motivate ourselves to do? Would it be pointless if it led the kleptomaniac to the conclusion that she has most reason not to steal? Perhaps the normal conclusion of deliberation would not have the *same* point that it has where one has no problem becoming motivated to do what one ought. If the kleptomaniac recognizes that she is unable or unlikely to follow the result of her deliberation, she will not yet know what to do. She will need to deliberate further about what she should do instead, such as avoiding circumstances in which she will be tempted to steal. But this does not make the initial deliberation or its initial conclusion pointless. Concluding that she ought not to steal supports her further deliberation and conclusion that she ought to find a way to avoid situations in which she might steal. In addition, for all we know, the conclusion that she ought not steal may have some effect over the long

term, even if it is not effective in the immediate instance.

There is nothing unusual about concluding that we really ought not to do something, then recognizing that we do not fully trust ourselves to avoid the temptation, and so coming up with a plan to avoid the temptation. It may be that I have most reason to go to a particular café to do my work. Suppose that it is within walking distance and therefore I could work there while loaning my car to a friend who would like to borrow it. But it may also true that if I tried to do so, I would probably end up wasting time using their free internet connection. In that case, I may come to the conclusion that I should keep my car and drive to my office instead, even while recognizing that this is a suboptimal solution. This does not make deliberating about what I have most reason to do pointless. Thus, Streumer's Argument from Deliberation may show that in order for deliberation to be useful our reasons must be limited in accordance with our abilities and opportunities, but it does not show that our reasons should be limited in accordance with our wills.

Therefore, I conclude that Streumer's arguments are consistent with the position that I have argued for here, and that, to the extent they are persuasive, they support my position.

CHAPTER TWO

GETTING REAL ABOUT MORAL RESPONSIBILITY

I. Introduction

In Chapter One, I argued that determinism would not undermine the truth of common sense normative claims about what people ought to do. However, discussions of determinism and morality often focus not on that issue, but instead on whether it makes sense to think of ourselves as morally responsible if determinism is true. It is natural to ask how it could be appropriate to blame, to resent, to shun, and to punish, or even to praise and reward, a person for an action that was determined by past events that were entirely beyond that person's control. These worries about moral responsibility are sometimes put in terms of a supposed inability to do otherwise: how could a person be morally responsible for an action if she could not have done anything else? But they need not be put this way. We might simply ask how a person could be responsible for an action that is the inevitable unfolding of events beyond the person's influence or control.

In this Chapter, to begin addressing these questions, I focus on some preliminary questions: what does moral responsibility involve and on what grounds might it be deserved? This chapter presents a descriptive account of blame and blameworthiness and of praise and praiseworthiness for actions and omissions. By a descriptive account, I mean one that is primarily intended to capture the essence of our actual attitudes or practices and our actual usage of terms like "blame" and "blameworthy." An account of blame and praise is, of course, not a full account of moral responsibility. We hold each other responsible in many ways. A full survey of these practices would be

long and unwieldy, and I suspect, an illustration of diminishing returns. The bulk of my focus shall be on blame, and as we shall see, it is difficult enough to get a grasp on that. This focus is appropriate. Blame is particularly important because, I shall argue below, it provides a foundation for some of the other ways we hold people responsible for bad behavior. In particular, if we can show that blame is justifiable, we will have gone some ways towards showing that retributive punishment could be justifiable as well, and if we cannot support blame and praise, then I see no way of supporting many of the other ways in which we hold people responsible for bad behavior.

II. Strawson's Argument for the Impossibility of Moral Responsibility

A clear account of blame, and of moral responsibility more generally, is essential to identifying the conditions that make blame and moral responsibility appropriate. As a result, such an account is also crucial in trying to understand whether determinism (or anything else) would make those conditions unsatisfiable.

To illustrate this point, consider Galen Strawson's argument for his position that "true moral responsibility" is "impossible." According to Strawson, "it makes no difference whether determinism is true or false. We cannot be truly or ultimately morally responsible for our actions in either case."[47] Strawson states his argument for this conclusion as follows:

> 1) Interested as we are in [action for which a person could be responsible], we are particularly interested in actions that are performed for a reason (as opposed to 'reflex' actions or mindlessly habitual actions).
>
> 2) When one acts for a reason, what one does is a function of how one is, mentally speaking. . . . .

[47] Strawson, Galen (1994) p.212.

3) So, if one is to be truly responsible for how one acts, one must be truly responsible for how one is, mentally speaking—at least in certain respects.

4) But to be truly responsible for how one is, mentally speaking, in certain respects, one must have brought it about that one is [that way]. And it is not merely that one must have caused oneself to be [that way]. One must have consciously and explicitly chosen to be [that way], and one must have succeeded in bringing it about that one is that way.

5) But one cannot really be said to choose, in a conscious, reasoned[] fashion, to be the way one is mentally speaking, in any respect at all, unless one already exists, mentally speaking, already equipped with some principles of choice, 'P1'—preferences, values, pro-attitudes, ideals—in the light of which one chooses how to be.

6) But then to be truly responsible, on account of having chosen to be the way one is, mentally speaking, in certain respects, one must be truly responsible for one's having the principles of choice P1 in the light of which one chose how to be.

7) But for this to be so one must have chosen P1[ ] in a reasoned, conscious, intentional fashion.

8) But for this, *i.e.* (7), to be so one must have already had some principles of choice P2, in the light of which one chose P1.

9) And so on. Here we are setting out on a regress that we cannot stop. True self-determination is impossible because it requires the actual completion of an infinite series of choices of principles of choice.

10) So, true moral responsibility is impossible, because it requires true self-determination, as noted in (3).[48]

In the initial steps of the argument, Strawson suggests that when one acts for reasons, one does so because of one's mental constitution, including, I would suppose, one's mood, values, beliefs, desires, and other mental traits and attitudes. This seems fair enough. If I steal when presented with an opportunity, it may be because I am greedy, or because I feel desperate and believe that I have no other options, or because I believe there is nothing wrong with stealing. The circumstances present me with (what I take to be) a reason for stealing only because I am "mentally speaking" one of these (or other similar) ways. Similarly, if I take a very slight provocation as a reason for lashing out, it could be because I am, "mentally speaking," irritable, short-tempered, and lacking in perspective. Others who are calm and even-keeled might disregard or be only mildly perturbed by a similarly slight provocation. So, I see nothing controversial in supposing that what we take to be reasons and which of those reasons we act on generally depends on what we believe, what we care about, how we feel, and more generally, our state of mind.

But consider the next steps in Strawson's argument, premises 3 and 4, which propose that in order to be responsible for action, one must be responsible for this underlying state of mind and, more specifically, that one must have consciously and explicitly chosen it. Strawson seems to be claiming that if I do lash out in response to slight provocation because I am irritable, I am responsible for doing so only if I "consciously and explicitly" chose to be irritable. These premises 3 and 4 are each substantive

[48] Strawson, Galen (1994) pp.212-13.

assumptions in Strawson's argument. They do not follow as inferences from premises 1 and 2. They are Strawson's assertions about the conditions of moral responsibility. Do we have any reason to accept them?

We can develop a case to test whether premises 3 and 4 are consistent with our intuitions about blameworthiness. Suppose that I am riding a crowded subway, and I am jostled and bumped by the other commuters. I become annoyed that my fellow commuters seem more concerned about clearing space for themselves than they are about jostling me. Suppose that on this particular hot and crowded day, I become annoyed enough to lash out at a fellow rider who I believe has been inconsiderate. Perhaps he stepped on my foot without apologizing and then continued to lean into me in an effort to give himself extra space. Rather than politely ask him to move, I lose my cool and give him a firm shove followed by an aggressive and angry stare. Assume that the situation defuses. Perhaps he is cooler-headed than I am, and soon afterwards one of us leaves the train, preventing any further interaction. As I calm down, I realize that the rider I shoved may not have meant to do anything wrong. He may not have even been aware of how his presence was imposing upon mine. In any case, my response was excessive and inappropriate. I also figure that the best explanation of why I lashed out is that I was unusually stressed and irritable as a result of things having nothing to do with him or with my commute. I took his conduct to be a reason for shoving him at the time only because of this stress and irritation, but I later realize that it was not a good reason at all.

As we are testing the plausibility of Strawson's assumptions, we need not yet accept the conclusion of his argument. So, assume for a moment, *contra* Strawson, that we are sometimes blameworthy. If we are, then the

foregoing incident seems to be a case in which I am. Shoving strangers and glaring at them for no good reason is not conduct I should be proud to tell others about. Instead, I should feel ashamed of this behavior, and I should seek to avoid behaving in this way in the future. In addition to blaming myself, I believe that I would deserve to be blamed and looked upon badly by those who witnessed or heard about my conduct. If I have the opportunity, I should apologize. Moreover, if I had happened to knock my fellow commuter over and cause some significant injury, I could rightly be held financially accountable, I might deserve some reasonable sort of criminal punishment, and I could not rightly complain if the victim or a bystander reported my assault to the authorities.

This poses a problem for Strawson's argument. In premise 4, Strawson assumes that in order to be blameworthy I must have "consciously and explicitly chosen" the state of mind that led me to act in this way. But my sense that I am blameworthy in my subway case does not depend at all on whether I ever chose, much less "consciously and explicitly" chose the stressed and irritable state of mind that led me to lash out in that way. Indeed, the very idea of such a choice can seem bizarre ("I have decided how I shall be: irritable and short-tempered.").[49] In my example, the act of lashing out might even be the first clear piece of evidence that makes me aware that I have, without ever thinking about it, become so irritable and easily provoked. If so, that might mitigate my blameworthiness, but it hardly absolves me of all responsibility. However I came to be irritable and short-tempered, the fact that

[49] Perhaps Strawson would claim that irritability is not the sort of mental state he has in mind. He might instead point to my belief that this fellow rider was being inconsiderate and that such inconsiderateness warrants aggressive retaliation. But these fleeting beliefs were certainly not consciously and explicitly chosen by me. They were the product of my irritation and stress.

I did not consciously and explicitly choose to be that way in no way prevents me from deserving blame and even punishment for what I do as a result.

There is nothing peculiar or unusual about the case I have described. Similar examples could be developed involving other common mental states that lead us to do things for which we may be blamed, including absentmindedness, distraction, jealousy, and many others. Few, if any of us, ever choose to be absentminded, distracted, or jealous, but we can deserve blame when these mental traits lead us to act badly. Thus, when applied to blameworthiness and blame, Galen Strawson's argument depends on assumptions about the conditions of moral responsibility, that are unmotivated and contrary to common sense intuitions about responsibility.[50]

The case I have described most clearly belies premise 4, but it also indicates that premise 3 is dubious as well. Without premise 4 to elucidate its meaning, premise 3 is problematically vague. If premise 3 is meant to say that

[50] One might counter by proposing a different premise 4: though I need not choose to be short-tempered, it must be that I *could have* chosen not to be so. This proposal helps Strawson's argument only if we can make sense of this counterfactual requirement and only if we can generate a similar regress by focusing on that counterfactual choice. But it is not clear that we can do either. First, any proponent of this response owes us an account of the sense of "could have chosen" at issue here. Identifying an adequate and plausible sense of "could have chosen" seems particularly difficult here. As suggested above, it is possible that I gradually became short-tempered without realizing it or ever meaning to become that way. Without awareness of this shift, it is not clear what kind of choice I ever really had about it. And even if we can identify some weak sense in which I did have a choice, it is not clear how the existence of that kind of choice helps explain why I am blameworthy. Second, the final steps of Strawson's argument seem to falter on this counterfactual approach. Strawson's argument generates a problematic regress because each of the choices required is a choice that must actually have been made. There is no objection that any one of the choices is impossible. It is only impossible for all of them to be made. And it is the fact that *each* must *actually* be made that means that *all* of them must *actually* be made. But on the counterfactual approach, at each step we would only require that a prior choice could have been made. Each of these counterfactual requirements may be satisfied: for principles of choice C1, I could have chosen not-C1; for principles of choice C2 relevant to that prior choice, I could have chosen not-C2; etc. We do not generate an impossible result if it is required that I could have chosen otherwise at each level. We generate an impossible result only if it must be that I could have made all of these other choices in the same counterfactual scenario. I cannot imagine any basis for this shift from each to all.

one must be morally responsible for the mental state that causes or explains the action, I take no position on its truth. In the cases I have described, I am uncertain but I suppose that I might be morally responsible for being irritable and short tempered even if I did not choose to be that way. But if premise 3 is meant to suggest that I must be causally responsible, that I must have affirmatively done something to bring about that mental state, then the case I have just described suggests that premise 3 is implausible as well.

So, why would Strawson think that premises 3 and 4 can bear any weight? Strawson's consistent use of "true" as a prefix to "moral responsibility" is a signal that his argument is targeted at a strong, and perhaps idiosyncratic, conception of responsibility. Strawson explains that conception here:

> What sort of 'true' moral responsibility is being said to be both impossible and widely believed in?
>
> . . . . As I understand it, true moral responsibility is responsibility of such a kind that, if we have it, then it *makes sense*, at least, to suppose that it could be just to punish some of us with (eternal) torment in hell and reward others with (eternal) bliss in heaven. . . . The story of heaven and hell is useful simply because it illustrates, in a particularly vivid way, the *kind* of absolute or ultimate accountability or responsibility that many have supposed themselves to have, and that many suppose themselves to have.[51]

Strawson's target matters. Though I am not prepared to endorse his argument as applied to heaven and hell, his problematic assumptions certainly become more plausible if we are talking about desert of heaven and hell. To see this, we need to consider the special significance of heaven and hell.

[51] Strawson, Galen (1994) p.216.

Eternal punishment and reward are not merely longer or more intense forms of responsibility than their human-imposed counterparts. They differ in kind. Human-imposed forms of responsibility come to an end. There comes a point at which the blameworthy or praiseworthy person has received her due. There may also come a point at which the person has shown through remorse or relevant changes in character, for example, that she no longer deserves to be held responsible for a past act. In contrast, the boundlessness of hell means that there is no point at which the wrongdoer will have properly paid for his sin. There is no point at which he will have suffered enough and the punishment should stop, and there is no point at which the wrongdoer will either deserve or receive a chance to redeem himself. Similarly, an eternity in heaven means that there is no point at which the person will have to show her moral worth again. As a result, heaven and hell have a kind of finality and conclusiveness that no form of human-imposed responsibility can have.[52]

Because eternal bliss and eternal torment are final and conclusive responses to human choices, they convey, and their propriety depends upon, a final and conclusive assessment of a person's "fundamental" moral worth. This is not a claim of religious doctrine, but of substantive moral theory. One cannot believe that a person deserves eternal torment unless one believes that the person is fundamentally bad and that there will be no need to revisit this assessment in the future or to provide that person with a new opportunity

[52] Capital punishment and imprisonment for life may seem to have a degree of finality. Without admitting that I am bound to defend these particular practices here, I would note that even these most final of human-imposed punishments come to an end. It is quite possible to suppose that, when they end, the person has been punished enough and has paid his due. Moreover, the lives of the people subjected to these punishments *can* always be reevaluated from a moral perspective (regardless of whether the law permits an opportunity for release or clemency). Finally, if we believe in an afterlife, we typically believe that their fate will be judged anew by a less fallible authority.

to prove whether he has changed. If one supposes that the assessment should be revisited at some point in the future, then one believes that the person deserves torment until that time, at which point he may or may not deserve further torment.

These considerations provide some support to Strawson's argument as directed at heaven and hell. My subway case suggests that we can be blameworthy for behavior caused by mental states that we did not choose. But this sort of blameworthy behavior does not, on its own, support condemnation to hell. An isolated instance in which irritability gets the better of me does not show that I am fundamentally a bad person who deserves eternal torment. My behavior might have been an aberration, the unusual result of an unlikely circumstance. The fact that I am susceptible to such behavior is significant, but it may be less indicative of who I truly am, than my normal, much more decent behavior or my nearly immediate feelings of remorse and shame. It could be that my short-temper and irritability are faults that I do not approve of and that I have been diligently striving to correct. Whether I am surprised at my behavior or not, the behavior may reflect traits that I am, in a sense, merely subject to, rather than revealing whether I am fundamentally good or bad. A final or conclusive assessment that could support eternal punishment requires behavior that reveals who I truly am, not behavior that reveals a disposition or desire that simply affects me or that I have not had an opportunity to master. Thus, my subway example does not appear to provide a very compelling counterexample to Strawson's argument when the argument is directed at heaven and hell. Instead, when dealing with desert of heaven and hell, the need to make a final and fundamental judgment pushes us to look for mental causes that are not merely given to us,

whether by deterministic cause or by indeterministic chance, but instead that have some deeper connection to our true character. Strawson's premises 3 and 4 arguably identify one way in which we might seek to assure such a connection. Choice is a natural proposal for establishing that the mental causes of our actions flow from and reflect who we truly are, rather than some merely given influences.[53]

This is not to say that I endorse Strawson's argument once it is directed at eternal torment and eternal bliss. It is not clear to me that choice is the only condition that could establish an adequate connection between the mental causes of our actions and our fundamental nature (assuming that there is such a thing as a fundamental nature). Nor is it clear that, if choice is an appropriate condition, it must be the sort of choice that generates Strawson's regress. I have to admit that my grip on, and interest in, the idea of a fundamental moral nature is not strong enough to resolve these issues. But I do find it to be clear that desert of heaven and hell would require that it be possible to make some sort of final and overall assessment of a person's moral worth, and that this in turn would need to be based on something more

[53] Note that I am not appealing to the intuition that a person must have an adequate opportunity to avoid hell. T.M. Scanlon appeals to this idea in explaining why Strawson's conditions may apply to hell, but not to his own account of blame. Scanlon (2008) pp.182-85, 190. The idea that a severe punishment, such as hell, should be avoidable has some intuitive appeal, but I do not find it adequate to explain the appeal of Strawson's argument, nor to show why it would not apply to responsibility more generally. The intuition does not explain why Strawson requires "conscious and explicit choice" rather than the mere opportunity to avoid. Nor does it explain why the argument would apply to heaven as well as hell. Most importantly, this intuition does not capture why hell might be categorically different than blame. Without some categorical difference of the sort that I have suggested above, we may be on a slippery slope. If, for example, the unpleasantness or harshness of hell supports the need for an adequate opportunity to avoid, then why does not the unpleasantness of normal punishment here in the real world? And if that does, then why does the unpleasantness of severe blame or the unpleasantness of mild blame?

than the sorts of behavior that can support the temporary and terminable blame that we impose here in this world.

This presents us with a challenge. I have argued that desert of heaven and hell require the propriety of a certain kind of judgment or assessment about a person, and that the nature of this assessment in turn is needed to support the conditions that Strawson imposes. But in rejecting the application of these same conditions to the desert of blame and punishment, I have relied only on intuitions about particular cases of blame. I have neither developed this idea of blame into an alternative to Strawson's understanding of moral responsibility, nor given any positive account of the sort of conditions that are required in order for a person to deserve to be held responsible in these ways. To get a grip on what those conditions are, we should first understand what blame and praise are. If heaven and hell represent "true" moral responsibility, I shall argue that blame and praise are the core elements of "real" moral responsibility, that which may be imposed by human persons here in this world. This is a more mundane topic than Strawson's, to be sure, but it is one of considerable importance nonetheless.

III. Proposals for an Account of Blameworthiness and Blame

Blame has been the subject of considerable philosophical attention in recent years. In this section, I consider a number of views of blame, including George Sher's dispositional account; a proposal, based on P.F. Strawson's work, that emphasizes the role of so-called reactive attitudes; and T.M. Scanlon's relationship-based account. I argue that these approaches do not reliably track our intuitions about particular cases of blame. Both Strawson and Scanlon point us in the direction of an important general feature of blame,

which is that blame involves a modification to a person's moral standing, but, I argue, neither develops this idea in the right way.

A. Blaming is Not Merely Judging or Stating that a Particular Person is at Fault.

There is a sense in which we blame merely by identifying a person as the person *to blame*, whether we do this in an unexpressed judgment or in a communicative act. "Finger pointing" is a common term for this sort of blaming. If there is a question about who committed some bad act or produced some bad consequence, and John says "James did it," then it will be true to say that John "blamed" James. To blame in this sense, John need not even believe that James was at fault. Of course, John's blaming act could be backed by a belief that James really is the person at fault. But in either case, John has pointed the finger at James, and thus, James could ask, "Why did you blame me?"

This is not the sort of blame that I am interested in. A simple example shows that there is a richer and more interesting kind of blame. Suppose that a colleague has done something that you find morally repugnant, and as a result, you are disappointed and angry with him. You may have distanced yourself from him to a certain extent, but your lives are intertwined in ways that make it difficult to cut off all ties. You must work together, and you have shared friends, commitments, and projects. Your disappointment and anger with him makes this difficult. He might say to you, "I know that what I did was wrong, and I understand why you are upset with me. I deserve it. But please, you need to stop blaming me. We need to be able to put this behind us if we are going to work together." This may or may not be a fair request. We can set that issue aside. The point is that when he asks you to stop blaming him,

he is not asking that you stop believing that he is at fault, or the person to blame. After all, he is admitting that he was wrong and that he deserved to be held responsible. Nor need he be asking you to stop identifying him as the person to blame. He may understand that you should continue to answer questions about who is to blame truthfully. Instead, what he is asking is that you put the issue in the past and stop letting his blameworthy acts affect the way you feel about him and the way you interact with him. In other words, by "blaming," he refers not to any judgment or speech act identifying him as the person at fault, but instead to the change that such a judgment has created in the way that you treat him.

There is another reason to suppose that blaming is something more than simply stating or judging that a particular person is at fault. Blameworthiness is the state of deserving blame. "A is blameworthy for X" means that A deserves blame for X. But saying or judging that A is blameworthy for X is a perfectly good way of saying or judging that A is at fault for X. So, if blame is simply an act or a judgment identifying the person to blame, then attributions of blameworthiness would seem somewhat empty. Admittedly, attributions of blameworthiness would not necessarily be meaningless. Attributions of blameworthiness might convey not only that one blames A, but also make it explicit that one believes it is appropriate to do so. However, when we say that a person is blameworthy, we presumably mean something more than that he deserves to be identified as the person to blame. We seem to mean that the person actually deserves to be held responsible in some way. Thus, we seem to be indicating that the "blame" of which the person is "worthy" is something more robust.

Some examples help confirm that we do maintain a distinction between judging a person to be blameworthy and actually blaming them. First, suppose that Raul reasonably judges that his boss is blameworthy for having an extra-marital affair, but that he also believes that it is none of his business to care about this. Not only might Raul think it is imprudent to concern himself with his boss's personal life, he might also feel that it would be inappropriately nosey and intrusive. And so Raul might decide that he will leave it to others to blame his boss, and that he will simply keep his relationship respectful, cordial, and professional. Thus, Raul can believe that his boss is blameworthy, while also believing that he is not the one to say or do anything about it.

As a second example, suppose that Raul is told about an episode in which George, someone he knows only by description, did something morally wrong without a decent excuse, but this episode has nothing to do with Raul or anyone he knows. If asked whether George is blameworthy, Raul might agree that he is. But if asked whether he blames George, Raul might find the question odd. He could reasonably respond, "Do I *blame* him? I don't know even know him. I don't condone what he did. There's no excuse for that sort of behavior, but it really isn't any of my concern."

Third, suppose that Raul's younger brother Daniel gets drunk and wrecks their father's car. Raul can admit that Daniel is fully blameworthy and that Daniel is going to deserve every bit of his father's anger. But Raul might say, "He's got it coming to him, and he's going to deserve it. But I have no right to blame him." He might say this because he feels he would be a hypocrite to blame, having driven the car drunk himself and having never apologized or repented for it.

These cases each suggest that a person may believe that another person is blameworthy without actually blaming that person in a more robust sense. This distinction between judging a person to be blameworthy and blaming that person would collapse if we thought of blame solely as judging or saying that a particular person is to blame. Thus, we have reason to look for an account of a more robust sort of blame.

B. Sher's Account(s) of Blame

George Sher agrees that blame is more than a mere judgment attributing fault. He approaches the topic of blame by asking "what blaming someone adds to believing that he has acted badly or is a bad person."[54] Thus, Sher appears to be after our same quarry. After considering a number of views, Sher offers the following proposal:

> The additional element . . . is a set of affective and behavioral dispositions, each of which can be traced to the single desire that the person in question not have performed his past bad act or not have his current bad character.[55]

Here, Sher seems to suggest that blame involves two elements in addition to believing that someone acted badly: a desire and certain kinds of dispositions.

With respect to the desire element, Sher suggests two possible contents: that the person not have performed the bad act and that the person not have his bad character. These desires are not meant to be interchangeable. The first sort of desire is involved when we blame a person for a bad act, the second sort when we blame a person for his bad character.

---

[54] Sher (2006) p.112.

[55] *Id.*

Sher argues at length that we can and do blame people for their traits and characters. As I am focusing on blame for bad acts or bad behavior, only the first of these desires is relevant.

Next, with respect to the dispositional element, Sher explains that the relevant dispositions include dispositions to feel negative emotions toward the blamee, such as "anger, resentment, irritation, bitterness, hostility, fury, rage, outrage . . ."; dispositions to engage in hostile behavior, which might take virtually any form, including "writing someone out of our will," "urinating in someone's flower bed" and, presumably, simply ignoring or giving the cold shoulder to that someone; and dispositions to reproach the blamee.[56] Sher claims that blame is closely associated with these attitudes and behaviors, but that a person can blame without actually having these attitudes and without actually engaging in hostile behavior. Sher therefore relies on dispositions in order to explain the connection between blame and these attitudes and behaviors, in a way that permits the existence of blame without them.

Though in the passage quoted above Sher seems to give dispositions a primary role, Sher is evasive about whether or not dispositions are actually essential to blame. Immediately, after that passage, Sher asks:

> But how much of this additional element is essential to blame and how much is contingently associated with it? To arrive at the best version of my proposal, should I take blame itself to encompass only the central desire-belief pair or also the various dispositions that that pair supports?[57]

[56] *See id.* pp.94-96.

[57] *Id.* p.112.

Sher is cagey about answering this question. He claims that "some affective and behavioral dispositions may indeed be essential to blame—but . . . augmentation [of an account of blame with the desire that explains those dispositions] renders its reference to these dispositions superfluous . . . ."[58] Sher also asserts that he need not provide a definitive answer to his question "because the focus of our inquiry is neither a word nor a concept but a phenomenon in the world[,] . . . [b]ecause my topic is the nature of blame itself—because I am seeking neither a conceptual analysis nor a dictionary definition but something more akin to a theory . . . ."[59] Further, Sher supposes that it does not matter whether we say that only the desire is essential or that the dispositions are essential as well, because, he claims, the blame-related dispositions are the "(virtually) universal" effect of the blame-essential desire, and that cases in which one has the desire without any of the dispositions are "so nonstandard" as to be "irrelevant."[60]

To be frank, I see no sense to the supposed distinction between conceptual analysis and theory that Sher is trying to make. Whatever the distinction is supposed to be, it cannot support Sher in avoiding the question of whether or not the dispositions are or are not essential to blame.[61] And for

---

[58] *Id.* p.98.

[59] *Id.* p.112.

[60] *Id.*

[61] The idea that one might provide an account of the phenomenon of blame apart from or instead of the concept of blame strikes me as deeply misguided. The contours of the phenomenon of blame are not simply out there to be found in the world apart from our concept of blame. We respond to bad behavior in all sorts of ways. There are no little markers out there in the world informing us which of those responses are blame and which are not. Of the possibly infinite varieties of responses to behavior, the phenomenon of blame includes those responses that share features we take to be interesting and important in some way, and excludes those that we do not. Thus, we cannot describe or give a theory of the phenomenon of blame without also specifying our concept of blame. In any case, the suggestion that one is providing a theory of a type of phenomenon invites the questions, "Which phenomenon is that, and what features does it have?" All other things being equal, a theory of a kind of

reasons that I shall explain below, I do not share Sher's belief that the desire he identifies reliably leads to the dispositions. Therefore, I will take Sher to have two different proposals:

> Sher1 - Blame for actions is a belief-desire pair, consisting of the belief that a person has acted badly paired with the desire that that person not have performed that past bad act.
>
> Sher2 - Blame for actions is a belief-desire pair of the sort identified in Sher1 together with dispositions to feel certain negative emotions toward that person, to engage in behavior hostile to that person, and/or to reproach that person.

Neither of these proposals provides an adequate theory of blame.

Let us begin with Sher1. This proposal features the idea of a past-oriented and now-unsatisfiable desire: the desire that a person had not performed a bad act. Sher supposes that, although unsatisfiable, this desire still has effects and, except perhaps in people with "nonstandard psychology," produces dispositions to negative emotions, hostile behavior, etc. Before considering that suggestion, we should pause to note that the very idea of a past-oriented desire can sound a bit odd. We do not normally attribute "desires" that the past were different. The words "desire" and "want" are typically used to describe a desire, preference, or yearning for something in the present or the future.[62] But in the broad sense of "desire" often used in philosophical discussions – a sense that covers a broad range of attitudes contrasted with beliefs – it may make sense to speak of past-oriented desires.

---

phenomenon is better to the extent it answers these questions with a robust and precise account of the features that instances of this phenomenon always, sometimes, or never have. Sher is simply refusing to answer these sorts of questions.

[62] We do often attribute desires that the present be different in some way that would require the past to have been different. For example, an orphan may *want* his parents to be alive. But this is not the same as specifically *wanting* or *desiring* that their death did not occur.

Such attitudes are just more commonly referred to as "wishes." In short, we *want* things to be different (now), but we *wish* that things had gone differently.

Characterizing the desire in Sher1 as a wish does not downgrade its significance or prevent it from playing the motivational and disposition-grounding roles that Sher attributes to it. An unsatisfied wish can be a very strong motivating or explanatory force. We often explain behavior by reference to wishes about the past. A retiree might take a "Great Books" course because he has long wished that he had gotten a liberal arts education. A person might lie about his past without expecting much benefit from it, primarily because he wishes that past were different. And of course, a person may be led to depression and all sorts of desperate acts because of the things he wishes he had done differently.

Using the more familiar term helps us to construct cases that will enable us to evaluate Sher1. If we are not familiar with cases in which a person has a desire that a person not have performed a past bad act, we should be quite familiar with cases in which we wish that a person had not. Those familiar cases show that (a) Sher1 is subject to counterexamples, because people very often have the belief-desire pair in question without blaming, and (b) that the belief-desire pair identified in Sher1 very often does not give rise to the dispositions that Sher takes to be characteristic of, if not essential to, blame.

Recall Raul, who does not blame his brother for wrecking their father's car because it would be hypocritical of him to do so. Raul recognizes that his brother acted badly. He also surely wishes that Daniel had not driven drunk and wrecked the car. He cares about Daniel and does not want to see him in trouble. He may also have been counting on borrowing that car himself. And we can also suppose that Raul wishes that Daniel had not wrecked the car

because it shows disrespect and causes hardship for his father.[63] So Raul can believe that Daniel did something wrong, wish that Daniel had not done it, and yet choose not to blame Daniel. He may feel that he should not blame Daniel, and so he may choose to support him and sympathize with him instead. As before, it still seems that Raul believes that Daniel is blameworthy but does not himself blame Daniel. The addition of Sher's wish or past-oriented desire does not change the situation.

The same may be true in the case of Raul and his philandering boss. He may judge that his boss acted badly, and he may wish his boss had not done so. Perhaps he feels badly for the spouse, perhaps he is saddened by the damage to their relationship, and perhaps he even feels badly for his boss, who may be suffering harsh emotional and personal consequences. Yet Raul may decide to stay out of it and to remain respectful, friendly, and even

---

[63] My list of Raul's reasons for caring calls attention to the fact that a person may wish that someone had not acted badly for reasons other than concern for the underlying moral considerations. Raul wishes Daniel had not acted badly in part because he does not want Daniel to suffer and because he may have wanted to borrow the car. The example does not *depend* on such non-moral reasons because Raul's respect and sympathy for the car's owner could also support the requisite backward-looking desire.

Nonetheless, the suggestion opens up a broad array of potential counterexamples to Sher1. For example, I may believe that you acted badly when you robbed that bank, and I may wish that you had not done so because now my own heist looks much less daring in comparison or because I had plans to rob that same bank the following day. And when I get back at you by turning you in, your defense attorney may also wish you had not acted badly because he had been planning a vacation. If we take Sher1 at face value, these are counterexamples. Sher1 calls only for a belief that a person acted badly and a desire "that the person in question not have performed his past bad act." Both your attorney and I agree that you acted badly, and we both wish you had not. But neither your attorney nor I *blame* you for acting badly. I am merely jealous, and your attorney is inconvenienced.

However, these probably are not counterexamples to the spirit of Sher's proposal. Sher describes the desire as "a specifically moral desire—because it is directed at the non-existence of a bad act . . . ." Sher (2006) p.109. For reasons we have just seen, that "because" clause is mistaken. A desire for the non-existence of a bad act is not necessarily a moral desire, even if the badness of the act is among the reasons for the desire. Nonetheless, I will assume that what Sher intends is that blame requires a specifically moral desire that the person not have performed that past bad act, *i.e.*, a desire that arises out of one's concern for the very moral considerations that make the act bad. Therefore, I am not relying on counterexamples that depend on clearly non-moral desires.

sympathetic to his boss. Thus, it would seem that Raul can have the belief-desire pair, but lack the corresponding dispositions, and it would be odd to say that Raul blames in this situation.

Sher1 also faces a serious problem in cases of forgiveness. Forgiving puts an end to blaming, but it need not put an end to believing that the person acted badly or to wishing that he had not done so. Returning to the drunk driving example, suppose now that there would be no hypocrisy in Raul blaming Daniel, but instead that Raul, recognizing that his brother is remorseful, forgives him and chooses to provide support. Raul recognizes that his brother acted badly, and he may strongly wish that his brother had not done so. Raul does not deny that Daniel is blameworthy, and he may believe that others would be well within their rights to blame or even punish Daniel, but for his own part, he has forgiven and does not blame. Furthermore, having forgiven Daniel, Raul is not disposed to be angry or hostile towards him. Raul has the belief-desire pair identified in Sher1, but he is not disposed to anger and hostility, and he does not blame.

Tragic figures also present a problem for Sher1. We can acknowledge that they are blameworthy, and we often sincerely wish that they had not acted badly, but we tend to feel sympathy and sadness rather than anger, hostility and blame. We can see Hamlet's tendency to brood and his failure to act swiftly as moral failings, but our belief that he is blameworthy may not lead us to *blame* Hamlet, and it certainly need not lead us to feel anger or hostility. Instead, we may feel sadness and sympathy. Or in Season 2 of the television series, *The Wire*, Ziggy is clearly blameworthy for his ill-fated decision to sell drugs and for the increasingly poor decisions he makes when this plan goes awry. It would be difficult to watch the series without wishing that he had not

done these things to himself or to others who are harmed as a result. But it is quite normal to feel only sadness and sympathy for Ziggy, not to have any disposition toward anger or hostility.

The absence of blame in these cases is not simply a form of detachment that comes with fiction. There are real tragic figures who generate a similar reaction. Many of them are likeable, brilliant, or, in many ways, admirable people who have significant flaws that lead them to frustrate their own interests and cause harm to others. We may recognize that those flaws are not going to change, and yet continue to care for and respect such people. Though such a person might be mentally troubled, he need not be, or at least he need not be so troubled as to preclude responsibility. We can believe that such people are accountable to those they harm and, if their acts are criminal, to the public at large. Yet, for our own part, we may accept that the flaws that cause these bad acts are simply a part of who this person is. Having accepted them, we may choose not to blame. We may recognize that their actions are bad, and sincerely wish that they had not performed them, but this belief and desire will tend only to produce sadness and sympathy, not anger or hostility, and not the sorts of changes that we would normally think of as blame.

Acceptance of flaws is not limited to tragic cases. We can accept all sorts of flaws. Consider a wife who comes to accept that her husband of many years has a temper or that he harbors certain unjustifiable prejudices. She need not approve of her husband's flaws or the resulting behavior. She might wish very much that her husband did not throw temper tantrums or did not make racist remarks. But having long ago recognized that these things are not going to change, she might no longer be susceptible to feelings of frustration, and she might have no disposition to be upset or angry. She has

learned to simply wait for the temper to pass or to ignore the comment and change the topic. Depending on the behavior and the circumstances, it could be wrong for her to accept such flaws. That is not my concern. What matters is that people sometimes do accept blameworthy actions in this way. When they do, they can believe that the person acted badly, and they can wish that he had not done so, but they are not disposed to feel anger or hostility, and they may eschew blame altogether.[64]

These examples should make clear that we can believe that someone acted badly, wish or "desire" that they had not done so, and not blame. It is at least as certain that we can, and very often do, have this belief-desire pair without having any of the dispositions that Sher claims are almost universally associated with blame. Therefore, Sher1 is inadequate as an account of blame.[65]

Let us turn to Sher2. If the dispositions that Sher identifies as relevant to blame are not entailed by the belief-desire pair in Sher1, as I have argued, then perhaps their addition to Sher1 will make for a more adequate account of

---

[64] By way of support for the view that the wife does not blame, consider the idealistic daughter who cannot understand why her mother tolerates her father's behavior. She blames her mother for not blaming her father.

[65] There may be another problem with Sher1. The belief-desire pair is directed at the wrong object to count as blame. Blame may be prompted by an event (an action or an omission), but it is directed towards an agent. We blame *for* a particular action or omission, but the *object* of blame is the agent, not the event. The belief and the desire that make up Sher1 are both directed at the event. Sher1 involves believing that a person acted badly and wishing that he had not. That is simply to believe that an event occurred and to wish that it had not. The fact that a person figures in the description of the event does not show that these attitudes are directed toward that person. It may be that these particular event-directed attitudes generally entail or are closely associated with some agent-directed attitudes. But it is not obvious what the agent-directed attitudes might be. As we have seen in the examples above, believing that a person acted badly and wishing that he had not are compatible with a variety of attitudes toward that person. The mere likelihood of an unspecified association with some agent-directed attitudes or other does not make this an adequate account of blame. An account of blame must identify and explain the stance that we take towards the person we are blaming.

blame. To understand Sher2, we need to say a little bit more about what dispositions are and why Sher takes them to be relevant to blame. Sher proposes an account based on dispositions because he believes that negative emotions and attitudes, hostile behavior, and reproach are commonly, but not always, associated with blame. Dispositions are supposed to explain blame's association with these attitudes and behaviors without making the attitudes and behaviors themselves necessary for blame. Dispositions can play this role because they are a sort of tendency or propensity to do something in certain circumstances. Dispositions tend to generate the specified behavior in the relevant circumstances, but they do not necessarily or always do so.

Of course, as with abilities, when discussing dispositions, the circumstances in which the specified behavior tends to occur are often not expressly stated. For familiar dispositions like fragility, we expect that common sense will provide at least a rough sense of the relevant circumstances and triggering conditions. But if we do not have a decent sense of what the triggering circumstances are, it will be difficult, at best, to determine whether an attribution of a disposition is correct. If I tell you, "That rock is disposed to shatter . . .", and you do not hear me continue ". . . at temperatures of negative 100 centigrade or colder," then you will be surprised when you cannot break the rock, and you may think me misinformed or a liar. The disposition to shatter when cooled below negative 100 centigrade is a very different property than the disposition to shatter at room temperature.

Sher does not describe the range of circumstances in which we should expect the blame-related dispositions to be triggered. This leaves Sher2 vague. Suppose that I insult someone. According to Sher2, should we say that she blames me if she is disposed to feel angry in circumstances in which

she is already tired and irritated by other things, or need she be disposed to feel anger no matter what? Need she be disposed to hostile behavior regardless of what else is going on, or is it enough if she is disposed to hostility only in circumstances where she is not distracted by some momentous news? It seems likely that most humans have standing dispositions to get angry, to engage in hostile behavior and to reproach in *some* circumstances. I, for one, currently have dispositions to do each of these things towards anyone in circumstances in which they do something infuriating. But these standing dispositions do not show that I currently blame anyone, even if they are coupled with the Sher1 belief-desire pair.

So, we should clarify Sher2 a bit. Presumably Sher intends that a person who blames is disposed to anger and to hostile behavior in a way that the non-blaming person who has these sorts of standing dispositions is not so disposed. I believe the best interpretation of Sher2 is that a person who blames must be disposed to anger and hostile behavior in precisely the circumstances she is in, those of having learned of some blameworthy behavior that she wishes had not occurred. If distraction means that the person is not currently so disposed, then distraction means that he does not blame. If frustration about other things helps generate the present disposition, then that frustration is part of the explanation of why the person blames.

Even with this clarification, Sher2 would leave us with an acute problem in determining whether or not a person actually blames or merely feels that a person is blameworthy. Sher2 is supposed to allow for cases in which a person blames, but does not actually feel anger and does not actually engage in hostile behavior. In such cases, it will be unclear whether a person's failure to feel anger and display hostility should be credited to the fact that he does

not blame or to the fact that dispositions do not always produce their specified behavior. That might be acceptable in cases where we are trying to determine whether someone else blames. We cannot, in fact, always tell whether someone blames. But this account of blame also makes self-knowledge problematic. As long as the dispositions do not actually produce anger, hostile behavior or reproaching behavior, the dispositions may remain indiscernible not only to others, but even to the person who supposedly blames. Now certainly a person can be confused or conflicted about how he feels towards another person. But the idea that he could actually blame someone without himself ever knowing it—without ever having any way of knowing it—strikes me as implausible.[66] Because others will be no better positioned to determine whether this person blames, no one will have any way of saying whether or not this person blames.

The problem is exacerbated by the fact that the Sher1 belief-desire pair does not reliably produce these dispositions, as I argued above in response to Sher1. We generally suppose that dispositions are grounded in other intrinsic or categorical properties of their objects. A vase is fragile, presumably, because of its structural and micro-structural properties; because, say, it is made of thin glass and because glass of that sort has certain molecular-level properties. If a person is particularly disposed to catch an illness, we presume that this has some underlying explanation, perhaps in terms of his genetic

[66] One could argue to the contrary. Imagine a person who becomes unduly angry at a friend for some minor misdeed. Casting about for an explanation he might find that his anger stems in part from an earlier and more significant trespass that had never been dealt with. The person might say, "I guess I did not realize it, but I have blamed you ever since then." I am doubtful that this would be correct though. Did he really blame his friend all along? Is it not better to say that he did not previously have a specific reaction, and that he is now blaming his friend? The fact that there has been some unaddressed irritation below the surface, waiting to be triggered by another event, does not convince me that the person has been blaming his friend all along.

makeup or his diet and sleep patterns. And if a person is disposed to walk the long way home through the park, we expect some underlying explanation of this. This explanation might be neurological or biological, but it could also be an explanation in terms of common sense psychology, *e.g.*, that he likes the scenery or that it is the only route he knows. A grounding explanation can be helpful in determining whether a thing has a particular disposition when its behavior does not make this clear. By looking for the grounding properties, we may be able to determine that a thing is fragile even if, through careful handling or simple luck, it has not shown any signs of breaking. So, if we had an account of the grounding properties for the dispositions in Sher2, we might at least have a criterion that, in principle, could distinguish cases of blame with untriggered dispositions from cases of non-blame. Unfortunately, Sher's suggestion is that these dispositions are grounded in the belief-desire pair identified in Sher1. He believes that this belief-desire pair generates these dispositions, except perhaps in cases of "imaginary people" with "nonstandard psychology." But as we have seen, there is a wide range of perfectly real cases in which the belief-desire pair do not lead to any of the supposed blame-related dispositions. Thus, we have no grounding account that might at least provide a theoretical criterion for blame.

These problems may not show that a dispositional account of blame cannot work, but they indicate that Sher has not given us an adequate one. They do show that Sher's account is untestable. In considering any given case in which a person does not actually become angry or hostile, we will have no basis for saying whether it is a case of blame or not.[67] The prospects

[67] Information about counterfactuals—*e.g.*, A would get angry with B, if B bragged about his bad behavior—is of little help. It does not demonstrate whether A had the relevant disposition in the actual scenario, whether he only had a weaker disposition to get angry upon

for an adequate dispositional account along these lines seem bleak. For one, dispositions alone, and thus in many cases un-triggered dispositions, add too little to the belief-desire pair to account for blame. If the belief-desire pair is inadequate for blame—indeed, if these attitudes are not even directed at the right object—it is hard to see how an un-triggered and potentially indiscernible disposition to have some attitudes could supply the difference. The person who has only the dispositions still lacks any particular attitude directed at the agent. If the dispositions remain un-triggered, there is nothing to suggest that his behavior and attitudes would in any way differ from those of a person who merely wishes that another had not acted badly.

Finally and most significantly, neither the proposed belief-desire pair nor the proposed behavioral and attitudinal dispositions account for the normative significance of blame, a feature that I will explain and argue for below. On Sher's account, negative attitudes and hostile behavior, where they occur, seem to be little more than outbursts, like a temper tantrum, stemming from a frustrated desire. But, I will argue, blame involves a change in our view of a person's moral status, a change in our view of the claims he may make against us and the expectations we may impose upon him. When we blame a person, we believe that he should feel ashamed or remorseful, we believe that he has little or no claim on our sympathy, and we believe that he ought to make amends for his behavior. Admittedly, these sorts of beliefs often accompany anger and hostile behavior. But the significance of blame lies not in those feelings, those hostile behaviors, or the dispositions thereto, but in what these attitudes and behaviors signify: a change in the blameworthy

---

some further provocation, or whether he had neither to begin with but the provocation in the counterfactual scenario generates a disposition to be angry.

person's moral standing. An account of blame that focuses solely on identifying the feelings and behaviors that are characteristic of blame, without reference to the normative attitudes underlying these behaviors, captures only the surface-level symptoms of blame. Therefore, we should turn to an account that does attempt to explain this normative component of blame.

C. A Strawsonian Account of Blame

Perhaps no writing has had a greater influence on philosophical views of blame (among broadly "analytic" philosophers at least) than P.F. Strawson's "Freedom and Resentment." Strawson has made it quite common to think of blame and moral responsibility in terms of "reactive attitudes," such as resentment, indignation and guilt. It bears emphasizing then that Strawson did not himself propose any account of blame. In fact, blame is hardly mentioned in "Freedom and Resentment." Strawson's goal in "Freedom and Resentment" was not to provide an account of blame, but to shift the focus of the free will and responsibility debate from what he characterized as "impersonal" reactions to textbook moral crimes, such as punishment, to the deeply personal and emotional reactions to grievances that we all experience, such as resentment. Strawson believed that this move raises the stakes for incompatibilists because it is harder to imagine that we could stop feeling these attitudes or that we would want to. Not everyone has shared Strawson's assessment of the possibility of either revising or giving up these reactive attitudes, but few of Strawson's readers would now deny that these reactive attitudes are an important part of how we respond to blameworthy behavior. Many have seen the possibility of extending them into an account of blame.

In developing a Strawsonian account of blame, two ideas in "Freedom and Resentment" merit attention. First, there are the reactive attitudes

themselves. Resentment and indignation are, arguably, ways of blaming a person, and guilt, shame, and remorse are plausibly important aspects of self-blame. Of these attitudes, Strawson suggests that indignation in particular is a distinctively moral attitude because it is based on concern for reasons generally. It is an emotional response to the sense that moral considerations have been flouted, whether that flouting affects oneself or some other person. As a result, a Strawsonian account of blame might propose that feelings of indignation, and perhaps the other reactive attitudes, are an element, if not the essence, of moral blame. Second, Strawson draws attention to what he sees as a consequence of holding some of these reactive attitudes, a consequence that he describes as a "withdrawal of goodwill":

> Indignation [and] disapprobation, like resentment, tend to inhibit or at least to limit our goodwill towards the object of these attitudes, tend to promote an at least partial and temporary withdrawal of goodwill . . . . The partial withdrawal of goodwill which these attitudes entail, the modification they entail of the general demand that another should, if possible, be spared suffering, is, rather, the consequence of continuing to view him as a member of the moral community.[68]

"Goodwill" is vague, but Strawson provides one concrete example of what it could mean to withdraw goodwill. In this passage, he suggests that, perhaps among other things, it involves a "modification" to the "general demand" that a person should be spared suffering. Thus, a withdrawal of goodwill may not mean merely having a less pleasant or friendly demeanor or having dispositions to anger and hostility towards a person, but seeing that person's moral standing as changed, and seeing them as having lost certain claims that

[68] Strawson, P.F. (1967) p.90 (emphasis removed).

they otherwise would have had. A Strawsonian account may take this moral change to be an important aspect of blame. So, a straightforwardly Strawsonian account of blame might hold that blaming involves (a) negative reactive attitudes, such as indignation, towards the person at fault, and (b) a partial withdrawal of goodwill, including a conviction that the person's rights or claims are limited.

Is such an account plausible? On the way to defending his own account of blame, Sher criticizes a Strawsonian account. Sher's treatment of the topic is a bit ham-handed in two respects. First, he incorrectly treats "Freedom and Resentment" as proposing an account of blame and inaccurately refers to "Strawson's account of blame." However, having pointed this out, we can generally ignore it. Second, Sher tends to blur the distinction between the two elements of a Strawsonian account. Sher initially describes the Strawsonian account as a "proposal that what blaming adds to believing that [a person] has acted badly is some sort of negative emotional reaction to him"[69] and then later as a proposal that "blame essentially involves anger or the withdrawal of good will."[70] Reconciling these descriptions, it would seem that Sher sees withdrawal of goodwill as a negative emotional reaction or a reactive attitude, rather than what Strawson says it is, a consequence of certain reactive attitudes.[71] But Sher describes his arguments as being directed at the view that blame involves either anger or lack of goodwill, so I will take them to be directed at any broadly Strawsonian

[69] Sher (2006) p.79.

[70] Sher (2006) p.85.

[71] Sher does not overlook the passages in which Strawson clearly states that a withdrawal of goodwill is an effect of the reactive attitudes and that it involves a modification of normative claims. He quotes them. See Sher (2006) p.79.

account, whether that account emphasizes the reactive attitudes, the withdrawal of goodwill, or both.

Sher offers two arguments against such accounts. Sher's first argument, strictly speaking, is not directed simply at a Strawsonian account, but at what Sher calls the "combined account," an account that combines the Strawsonian account with the view that blame can be deserved. As I am committed to this latter part of the "combined account," I, like Sher, would take any problems with the combined account to be a problem for the Strawsonian approach. Sher claims that the combined account entails that:

> it is appropriate for everyone to be angry at, or to lack good will toward, virtually everyone else he knows. That conclusion, however, is in effect a *reductio*; for whatever else is true, perpetual mutual hostility is hardly an appropriate ideal of human interaction.[72]

The starting premises and the conclusion of this argument alone should make one suspicious. The idea that the appropriateness of "perpetual mutual hostility" can be drawn out of a Strawsonian account of blame, like an elephant from a mouse hole, is a bit too much to believe. Not surprisingly, the argument that is supposed to accomplish this feat is deeply flawed.

The first steps in Sher's argument are stated as follows (with emphasis added to illustrate what I take to be the central premises and the conclusion):

> [1] We can, if we like, conceal any anger or lack of good will that we harbor toward a wrongdoer. Thus, [2] if it is true both that anger or lack of good will is what blaming someone adds to believing that he has acted badly and that blame as so construed can be deserved, then it must also be true that wrongdoers can receive all the blame they deserve

[72] Sher (2006) p.87.

> without ever knowing it. Even when someone receives his full measure of deserved blame, his receiving it need not affect his life. This implication is noteworthy because [3] in many other contexts a person's getting what he deserves necessarily does have some impact on his life. However, what the implication shows is not that the combined account is unacceptable, but only that in order to accept it, [4] we must take "X deserves blame" to mean no more than that blame directed at X is justified or appropriate—an interpretation that is in any case forced upon us by our earlier observation that *blame itself* can be kept strictly private.[73]

This portion of the argument is supposed to show that, because blame on the Strawsonian account can be concealed we cannot think of blame as being, strictly speaking, *deserved* by the blameworthy person. Instead, we must think of it as being something that is *appropriate* for the blamer to do. This is only a subconclusion or first step in Sher's argument, but we may stop here and consider the premises that are supposed to get us to this subconclusion. I accept [1], that blame on a Strawsonian account can be concealed from the target of blame (except perhaps in cases of self-blame). So, I will focus on the next two claims, which are [2] that this means that, on a Strawsonian account, a person can "receive all the blame they deserve without ever knowing it" and without it "affect[ing] his life," and [3] that "in many other contexts" getting what one deserves "necessarily" has some impact on one's life. Both of these premises are plainly false.

First of all, premise 2 should be rejected because nothing in either premise 1 or a Strawsonian account of blame supports, much less entails, that a person can get their full measure of deserved blame without it affecting his life. Sher fails to explain why it follows from the fact that *any* potential blamer

[73] *Id.* p.86.

can hide blame (premise 1), that the blamee might get *all* the blame he deserves in this hidden form.[74] Why shouldn't the Strawsonian say that, although as a factual matter any particular blamer can conceal his blaming reaction, as a moral matter blameworthy people generally deserve at least some unconcealed blame? Even assuming that it sometimes happens that every blamer conceals his blame for a particular person, why couldn't the Strawsonian quite plausibly respond by saying that sometimes people do not get what they deserve? The existence of concealed and unconcealed blame does not entail that people may receive all the blame they deserve solely in the form of concealed blame any more than the existence of closed-mouthed and open-mouthed kisses means that people may receive all the kisses they deserve solely in the closed-mouthed form. Just as a Strawsonian can believe that people generally deserve a few open-mouthed kisses from time to time, a Strawsonian surely can believe that blameworthy people generally deserve at least some unconcealed blame.

Even if the Strawsonian wanted to concede that a person might get all the blame they deserve in the form of concealed blame, there is a further problem. Premise 2 depends on an assumption that something affects or "has an impact on" a person's life only if that person is aware of it. That is simply wrong. My life would be terribly worse if, unbeknownst to me, the people I take to be my friends and associates actually harbor concealed negative attitudes or secretly lack goodwill towards me. I am much better off if my friends *actually* feel warmly towards me than if they merely pretend to. In

[74] It is also puzzling that Sher's argument is supposed to follow from premise 1—the idea that blame can be hidden—when, as explained above, blame can certainly be hidden on either of Sher's accounts. It is not clear why Sher's argument would not apply with equal force to Sher's account, if it had any force.

some cases ignorance might be bliss (*e.g.*, when knowledge causes only unnecessary and unproductive anxiety), but in this instance, true wellbeing involves correspondence between belief and reality.

Moreover, concealed negative attitudes or lack of goodwill can have concrete, negative effects on our lives without us being aware of either the attitudes or their effects. Concealed negative attitudes are quite likely to result in the closing off of opportunities that otherwise would have been open. But for her concealed negative feelings towards me, an acquaintance might have been inclined to become a closer friend to me, to go out of her way to help me, to invite me to a wonderful event, or to speak kindly of me to a third person. When, in fact, she fails to do these things, I may not realize that I am missing out on opportunities I otherwise would have had, much less suspect the reason. After all, I do not expect all of my acquaintances to go out of their way for me in these ways. Nonetheless, I am worse off than I would have been if she did not have these attitudes towards me. I have lost out on opportunities for help and friendship that I otherwise would have had. Thus, concealed negative attitudes can have a dramatic impact on our lives without our being aware of it.

Sher's next premise, that "in many other contexts a person's getting what he deserves necessarily does have some impact on his life," fares no better. Sher does not identify or describe a single one of these "many other contexts." But if he supposes, as he must for his argument to be valid, that a person's getting what he deserves "necessarily" involves that person being *aware* of an impact on his life, then he is wrong. There are contexts in which a person gets what he or she deserves without ever being aware of it. People who receive posthumous honors are very often said to be getting something

they deserve. The same is true of those who become exposed as traitors, criminals or scoundrels after their death. A former head of state, now senile and unaware, gets what he deserves when his crimes are exposed, he is convicted, and he is stripped of all of his honors and national recognition, even though he may live out the remainder of his life unaware that anything has changed. A prisoner of war or an explorer, entirely cut off from the rest of the world, might receive a deserved honor back in his home country.[75] There are also examples that do not depend on unusual ignorance-causing circumstances like senility or separation from society. A bad husband might be getting just what he deserves when, unbeknownst to him, his spouse is unfaithful to him, and a pompous jerk might be getting what he deserves when people make fun of him behind his back.

In short, nothing about a Strawsonian account of blame makes it problematic to think of blame as deserved by the blamee. We need not know that people blame us in order to be getting what we deserve. Thus, a defender of a broadly Strawsonian account need not be forced from the position of conceiving of blame as something that is deserved by the blamee to the position that blame is only something that may be appropriate for a blamer to engage in. Nonetheless, let us briefly consider whether the next part of Sher's argument would follow if we did concede this.

Sher claims that, as a result of the shift from desert to appropriateness, the Strawsonian will be unable to limit the appropriateness of blaming to a reasonable class of blamers. Instead, Sher claims, the Strawsonian is forced to conclude that it is appropriate for virtually anyone to blame anyone who has ever done anything wrong.

[75] Thanks to Paul Kelleher for providing or planting the seed for these examples.

> [1] Because the normative element of the combined account must thus be oriented to the person doing the blaming rather than the person blamed—because the account must view deserved blame not as something a wrongdoer "has coming to him" but rather as a response that his transgression renders appropriate in others—the account raises questions both about which others may appropriately have the response and about how long they may have it. . . .
>
> [I]f we are to capture the detached, impartial quality that Strawson rightly takes to distinguish blame from resentment,[76] [2] we can hardly restrict those in whom anger or a lack of good will toward a wrongdoer is appropriate to persons who are themselves affected by his transgression. There is, moreover, no non-arbitrary basis for excluding any people who are not affected. [3] For this reason, the combined view is best taken to assert that *anyone* who knows what a wrongdoer has done—including the wrongdoer himself—may appropriately react to him with anger or a withdrawal of good will.[77]

Sher believes that this entails the conclusion of his supposed *reductio*:

> Because we all sometimes act wrongly in ways that are readily apparent, it is not unreasonable to suppose that virtually everyone is known by each of his acquaintances to have recently performed at least one blameworthy act. This supposition . . . supports the conclusion that it is appropriate for everyone to be angry at, or to lack good will toward, virtually everyone else he knows.[78]

[76] Sher seems to be referring to Strawson's distinction between resentment and indignation, not resentment and blame. Strawson does not draw any distinction between resentment and blame in "Freedom and Resentment."

[77] Sher (2006) p.86.

[78] *Id.* p.87.

This phase of Sher's argument is at least as problematic as the first. The initial claim is baffling. It is not clear why or how Sher supposes that his reinterpretation of "X deserves blame" in terms of the appropriateness of blame raises any new questions about who may blame. On its surface, "X deserves blame" is just as unlimited with respect to potential blamers as "It would be appropriate to blame X." So, why can't Sher's reasoning be used to show that a person who thinks of blame in terms of desert is also committed to the view that a blameworthy person deserves to be blamed by anyone? If—as is actually the case—there are reasons why a blameworthy person does not deserve blame from anyone and everyone, why do those same reasons not also show that it is not appropriate for anyone and everyone to blame X? Put another way, when we say "X deserves blame," we surely mean only that "X deserves blame [from some appropriate blamers, but not everyone]." We do not mean that X deserves blame from unrepentant hypocrites; from people who have an overriding obligation to remain neutral, such as a counselor or a peer in group-therapy; or from those who do not know enough about the situation to properly calibrate their blaming response or to determine when X is no longer blameworthy. So, if we did accept the need for a translation of "X deserves blame [from some appropriate blamers, but not everyone]," what prevents us from insisting that that translation should be "it would be appropriate [for some appropriate blamers, but not everyone] to blame X"?

Sher's only stated support for his view that we must give up on an "appropriate blamers" restriction (which we might call a "standing" requirement for blaming) is his claim that there is "no non-arbitrary basis for excluding any people who are not affected" by the blameworthy behavior. This unsupported assertion is incredible. There are numerous non-arbitrary reasons for limiting

the universe of appropriate blamers, many of which exclude people who are not affected. I have just mentioned a number of categories of inappropriate blamers: hypocrites, those obliged to remain neutral, and those who lack adequate knowledge. Let me elaborate on this last category, as it is based on considerations that rule out many of the people who are not directly affected by the behavior prompting the blame. People who are more remote from the blameworthy person or act may come to learn of the blameworthy act, but by and large, they tend to be in a worse position to blame because they will, by and large, have less knowledge of mitigating factors, of the actual degree of harm suffered by any victims, of the blameworthy person's subsequent behavior, and of the countless other complexities that affect the appropriate response. Their blaming response is therefore apt to be excessive, and that in itself is a good reason for them not to blame, even though it may be perfectly appropriate for others to blame. It is also relevant that people who are far removed from the blameworthy person often have better things to worry about. Though we all have some reason to care about wrongdoing wherever it happens, we have only so much time, energy, and capacity to care. If, as Sher's conclusion supposes, people were constantly investing attention and energy into all of the routine moral failings of all sorts of people to whom they have no real connection, they would very likely have too little invested in those who are close to them. Even if they could manage some constant concern about all other distant people (and surely they could not), it would be *wrong* for them to do so at the expense of concern for others with whom they are more closely related. Thus, it can be quite inappropriate for many people to blame those who are blameworthy.[79] In sum, a Strawsonian account of blame is

[79] Sher's additional claim that the defender of an attitudinal account lacks a basis for

perfectly capable of accommodating the view that blame is deserved and the fact that, when blame is deserved, there may be considerations making it inappropriate for some particular people to blame.

Let us turn to Sher's second argument against Strawsonian accounts:

> The basic difficulty with that account is that the generalization upon which it rests—that blame is always accompanied by hostility or a withdrawal of good will—is far from airtight. . . . We may, for example, feel no hostility toward the loved one whom we blame for failing to tell a sensitive acquaintance about a hard truth, the criminal whom we blame for a burglary we read about in the newspaper, or the historical figure whom we blame for the misdeeds he performed long ago. As the latter examples suggest, blaming is something that we can do regretfully or dispassionately and that need not be accompanied by any rancor or withdrawal of good will.[80]

These supposed counterexamples are a bit of a mixed bag. Without some further development, it is difficult to see why we should think that the second and third in particular (the criminal in the newspaper and the historical figure) present cases of blame. They are quite like my example above, in which Raul

---

limiting the time period over which blame remains appropriate does not merit much additional discussion. Sher states that "to capture the important idea that a person can deserve blame now for something he did a long time ago, the combined view is best taken to assert that those reactions remain appropriate long after the transgression itself." Sher (2006) p.86. This is a simple error of taking evidence of *sometimes* to support *always*. It is true that sometimes a person can deserve blame long after her transgression. Some people are remorseless, some people evade punishment and responsibility for a long time, and some transgressions are so serious that they are not quickly paid for or forgiven. But this hardly supports the conclusion that, as a general matter, it is appropriate to blame people long after their transgressions. Very often people receive all the blame and punishment they are due; they apologize, show remorse, and make amends; and they change in ways that can render blame inappropriate. There is no reason why a Strawsonian account cannot incorporate these simple facts. Indeed it would be plausible for a Strawsonian (or anyone else) to assert that over time there is a growing presumption that blame is inappropriate, absent special circumstances of the sorts just identified above.

[80] Sher (2006) p.88.

is told about a blameworthy act performed by someone he has never met that has no direct impact on anyone he knows or cares about. As I said there, it might be normal to suppose that we disapprove of the conduct at issue and judge the person to be blameworthy, but it would be odd to suppose that we blame these complete strangers.

Of cases like this, T.M. Scanlon plausibly says:

> [Blame's] content is attenuated in the case of agents who lived long ago and have no significance for or effect on our lives. We can judge such people to be blameworthy, but such a judgment has mainly vicarious significance . . . . It may imply, for example, that those who interacted with this person had good reason to withdraw their intentions to trust or rely upon him. But the idea that we ourselves *blame* him for what he did can sound somewhat odd.[81]

It is true, of course, that we can sometimes have strong emotional reactions to stories of distant blameworthy activity. We can be disturbed, disgusted, and even angered, but of course, Sher is not talking about such cases. He is trying to identify a case of dispassionate blaming. It is not clear whether being moved to anger by a story of some distant blameworthy act should count as blame, but *not* being moved by a story of distant blameworthy behavior is clearly not a case of blaming. Without some further argument, I take Sher's history and newspaper examples to be what they seem, examples of judging a person to be blameworthy but not actually blaming.

Sher's other example, blaming a loved one without hostility, and his suggestion that we can blame "regretfully" are a bit more interesting. I think Sher may be correct here. After all, one of the more familiar ways of

[81] Scanlon (2008) p.146.

communicating blame to a person that we trusted, liked, or cared about is by stating that the person has disappointed us. In such cases, we may blame, but more with a sense of sadness than anger. This is not surprising. If we trusted or cared for the person we blame, it may be sad and painful to accept that our hopes and expectations of him have not been fulfilled or that we cannot have the relationship we had hoped, but at the same time, our affection for that person may keep us from feeling outright anger.

But how much does this show? For one, it is not clear that a Strawsonian account must be saddled with Sher's assumption that the relevant reactive attitude is anger, or any other attitude of hostility. Strawson did not suggest this, and it is not clear why we should. Why not suppose that a feeling of morally-based disappointment is a blaming reactive attitude? In any case, the mere suggestion that we can blame regretfully does not cast any doubt on the second aspect of a Strawsonian account, that blame involves a withdrawal of goodwill. As Strawson suggests, a withdrawal of goodwill need not be seen as a weak analogue for anger but, at least in part, a change in our view of the blameworthy person's moral claims. I see no reason in Sher's cases to believe that we can blame a person without feeling that the person has, to some degree or in some respect, less of a claim to sympathy, concern and respect.

Thus, neither of Sher's criticisms poses any obstacle to the development of a broadly Strawsonian account of any variety. First, Sher fails to demonstrate that we can blame entirely dispassionately. His point that we can blame regretfully does not rule out the possibility of a reactive attitude based account. It merely suggests that if we suppose that blame involves reactive attitudes, we should take a broad view of what those attitudes are.

Second, none of his criticisms casts any doubt on the suggestion that blame involves a modification to one's view of the blameworthy person's moral standing, including seeing that person's claims as diminished and, I would add, imposing new demands and expectations on that person. If there is a problem with a Strawsonian account of this sort, it is that it needs further development.

D. Scanlon's Relationship-Based Account of Blame

T. M. Scanlon's account of blame provides one way of developing the Strawsonian idea that blame involves a withdrawal of goodwill. Scanlon's account of blame is built around the idea of relationships. He suggests that ideal relationships involve persons having certain attitudes, intentions, dispositions and expectations toward each other. For example, on Scanlon's view, being friends ideally involves, among other things, "intending to give help and support when needed, beyond what one would be obligated to do for just anyone; intending to confide in the person and to keep his or her confidences in return; intending to spend time together when one can . . ." and presumably, liking and caring about the person.[82]

Scanlon claims that blame is a response to behavior that shows what he calls an "impairment" to a relationship.[83] What he means, it seems, is that we blame in response to behavior that shows a person does not have the attitudes, intentions, etc., that make up the ideal relationship, or at least the relationship we reasonably hoped or expected we had with that person. So, in the case of friends, one might blame if a supposed friend shows that she lacks the requisite affection, if she fails to keep confidences, or if she stops showing

[82] Scanlon (2008) p.132.

[83] *Id.* p.128.

interest in spending time together. Blame itself, on Scanlon's view, is a responsive modification to the relationship that is appropriate in light of the other person's lack of the requisite attitudes, *etc*.[84] In seeing the friend as blameworthy, we may believe that he has lost his claim to reciprocal friendship. So, blame in a friendship might involve a corresponding decrease in one's affection and concern for the other, an end to one's own intention to spend time together, or I would suppose, an intention to confront and reproach the other person.

Like the Strawsonian idea of a withdrawal of goodwill, blame on Scanlon's account involves taking a new or special stance towards the blameworthy person, including a revision to one's view of the claims that the blameworthy person has against oneself. But Scanlon's framework suggests a possible path for replacing the vague Strawsonian idea of a withdrawal of goodwill with something more specific. In theory, we might be able to identify the specific attitudes and expectations that are appropriate in a given relationship and the ways in which we modify those attitudes when we blame.

There are, however, at least two problems with this account. The first is structural. The relationship framework of Scanlon's account is strained in a variety of cases. Most obviously, it is strained in cases of pure moral blame among strangers, where there is no preexisting relationship. But even in cases involving a real relationship, we can often reasonably blame in response to behavior that does not reveal anything that could be naturally described as "impairment" to a relationship. Applying Scanlon's relationship framework in such cases is an imposition upon, rather than illumination of the underlying

[84] *Id.* p.128 ("To *blame* a person is to judge him or her to be blameworthy and to take your relationship with him or her to be modified in a way that this judgment of impaired relations holds to be appropriate.")

phenomenon. The second problem has to do with the content of Scanlon's view. Scanlon's own account of appropriate blaming responses is unjustifiably revisionist. In general, Scanlon is too mild or timid about the special attitudes that a person may take when she blames another. If we were to defend only Scanlon's version of blame against determinism and other worries, we would be conceding a great deal of our actual beliefs and practices. I will explain each of these problems in greater detail.

1. The relationship framework of Scanlon's account

The structure of Scanlon's account is strained when applied to cases of purely moral blame, where one person is wronged by a stranger with whom he does not have a preexisting relationship, in any normal sense of the word. Imagine, for example, a person who is mugged while traveling thousands of miles from home in a place where she knows no one. Scanlon acknowledges that it "sounds odd" to talk about a preexisting "moral relationship" in such cases and to propose that blame is a reaction to the impairment of that relationship.

> [W]e naturally take the term 'relationship' to refer to a *particular* relationship, like the friendship between two individuals, which is constituted by the friends' special attitudes toward each other. Morality is not a relationship in this sense. Rather it is a normative ideal . . . that specifies attitudes and expectations that we should have regarding one another whenever certain conditions are fulfilled.
>
> * * *
>
> In the case of morality . . . the relevant conditions [for the existence of the relationship] do not concern the parties' existing attitudes toward one another but only certain general facts about them, namely that they are beings of a kind that are capable of understanding and responding to

> reasons. Insofar as one assumes that the relationship must, like friendship, be constituted by the parties' attitudes, this provides a . . . reason for thinking it inappropriate to say that morality defines a relationship that holds even between total strangers. But this assumption is mistaken. The conditions in virtue of which the relationship exists, and the relevant normative standards therefore apply, do not always involve the parties' attitudes toward one another.[85]

This passage shows that there is *nothing* to the "moral relationship" except (a) the fact that the parties are persons capable of responding to reasons and (b) the existence of moral norms and moral reasons that apply to the way such persons should feel and act towards each other. Thus, the parties to the "moral relationship" need not have any actual attitudes towards one another. Indeed, they need not even know the other exists. So, Scanlon's "moral relationship" is simply an unusual way of referring to the fact that there are moral norms dictating how competent agents ought to feel and behave towards each other.

If we can describe the existence of norms and the capacity to respond to them as a "relationship," what does it mean to speak of "impairment" to that "relationship" as the thing that warrants blame? As there is nothing to the relationship except for the fact that people ought to be responsive to normative reasons, behavior revealing an "impairment" to this relationship could be nothing other than behavior showing that a person is not appropriately concerned about these norms.

My concern here is not so much that this terminology is wrong, but that it is unnatural and unhelpful. I suppose that we might agree to use terms like

[85] Scanlon (2008) p.139 (emphasis added).

“relationship” and “impairment” in these ways, but doing so obscures rather than illuminates the underlying phenomenon. The idea that my moral relationship with someone is impaired is an utterly foreign way of capturing the entirely familiar idea that they have failed to show the sort of concern for moral norms that I should expect. Thus, the explanation of what Scanlon means by his terms is clearer and more helpful than the account using these terms. As a result, at least in the case of pure moral blame among people with no prior relationship, Scanlon appears to be imposing a theory on the facts, rather than providing a theory that explains them.

Scanlon could concede that the relationship structure is strained in the purely moral case, but argue that its value lies not in its ability to describe this one kind of case, but in its ability to illuminate the differing nature of blame across various circumstances. Perhaps the idea of a moral relationship, though imperfect, provides a useful contrast with blame in cases of a preexisting relationship, and perhaps the ability to apply a consistent framework allows us to see why blame varies in these different contexts.

But even in the case of preexisting relationships, the idea that blame is a response to “impairment” of the relationship is a dubious or, at best, strained description of the phenomenon. Consider two brothers, the younger of whom irresponsibly runs out of money. The older brother may feel obliged to help with a loan, but because he blames the younger brother for his wasteful spending, he also feels that the younger brother has little claim to feelings of sympathy regarding his predicament. He may also feel that the younger brother deserves a stern lecture that, in other circumstances, would be condescending and inappropriate. Though he bails his brother out, he may, for a while, be irritated or even indignant about it, and he may let his brother

know how he feels. The older brother may also expect, or even demand, contrition and a commitment to be more responsible. A simple and seemingly adequate explanation of what is going on here is that the older brother blames the younger brother for not exercising care and good financial sense. Because it is his brother, he remains more helpful and more concerned than he would toward a stranger, and he still feels affection and love for his brother, but for a time, he is irritated and sees his brother as less trustworthy. He may see his brother as having forfeited a claim to trust in financial or other related matters and as being subject to new demands in order to earn that trust back.

On Scanlon's account, to see the older brother as blaming we must suppose that he is reacting to an attitude that shows some impairment either to their relationship or to the younger brother's relationship with others. But what relationship is impaired here? And is any such supposed impairment really a better explanation of what the older brother is responding to? It might be argued that there is an impairment to the relationship between the brothers. The younger brother has failed to get by independently and handle his own affairs. He has therefore shown some dependence on his brother and some lack of reliability. As a result, the older brother may feel that he cannot count on his brother to handle certain situations involving money, or for that matter, to be able to provide the same sort of financial support if their roles were reversed. But is this really an impairment to their relationship?

I certainly admit that we all have reason to be responsible for ourselves and not to be an unnecessary burden on others. In particular, we may owe it to our family to avoid being a burden as best we can. We may also owe it to our family to take care of ourselves, not only because we might burden them with our messes, but because our failures and our suffering may be painful for

them. But I do not accept that a failure to live up to these responsibilities—even a failure that merits blame—should necessarily be seen as an impairment to our relationship with our family or that any such impairment provides the best explanation of the blaming reactions. Indeed, I find it hard to understand the suggestion that an occasional failure to be self-reliant and financially responsible should count as an "impairment" to one's relationship to one's immediate family. Instead, the suggestion that a brother's financial irresponsibility deserves blame because it makes him a "bad brother" or because it damages the relationship between the brothers seems to me to be an insult to the ideal of brotherhood.

In my view, complete independence is not necessarily coincident with ideal family relationships. Instead, it seems to be characteristic of strong, close family relationships that their participants expect to take care of each other from time to time. Ironically, the younger brother's blameworthy mistake could actually strengthen their brotherly relationship. It could be that, having been independent, the brothers were drifting apart, and that the younger brother's need for financial help in my example reaffirms, rather than impairs their relationship. In that case, it would be very odd to explain the older brother's blame as a response to the younger brother's impairment of their previously weak relationship. Nonetheless, blame may be appropriate.

Nor is it plausible to suppose that the older brother's blame is a response to some impairment to the younger brother's relationship with others. We may assume that the younger brother's conduct will have little significant impact on the rest of the family's wealth, on their time and energy, or on their love and concern for the younger brother. We need not suppose that the brother has failed to pay any debts to others. Thus, at most, his

irresponsibility and dependence might show that he is not an ideal candidate for certain kinds of relationships going forward. Creditors may want to beware. But surely this is not why the older brother blames his younger brother. The older brother need not be concerned about the younger brother's relationships with potential, future creditors. He may blame simply because he is upset about what the younger brother has done.

Scanlon could respond that there is impairment to their relationship as brothers because, although they may expect to support each other when necessary, ideally they also have hopes and desires that each other is responsible and does well, and the younger brother has frustrated those hopes and desires. But that is simply to say that the younger brother has not done what he should and the older brother cares about it. Thus, in this case, the simpler explanation is the better one: in blaming, the older brother is responding to the fact that his younger brother has been irresponsible, or, in other words, not adequately concerned about acting on reasons and doing the right thing. What this case suggests is that blame is not in any normal sense a response to impairment of a relationship. It is a response to a person's failure to show adequate concern for reasons.

This is not to say that relationships are irrelevant in assessing or attributing blame. The reasons that a person has depend upon his or her relationships. We have reason to show basic concern and respect for people generally, but we also have reasons to do other things that arise out of our friendships, our family ties, our group affiliations, our contractual relationships, etc. Relationships may also affect our interest in the blameworthy act and our standing to blame, as well as the form that our blame should take. So, we must be attentive to the existence of various relationships and the reasons

that arise from them in order to determine when a person will be blameworthy and from whom he will deserve blame, but this does not mean that it is accurate or illuminating to think of blameworthy behavior as behavior that reveals impairment to a relationship or to think of blame as a reciprocal response to that impairment.

2. The anti-retributive aspect of Scanlon's account

Though the structure of Scanlon's account is problematic, his suggestions about the particular reactions that constitute blame still deserve consideration. Scanlon identifies a handful of special attitudes and dispositions that may arise in cases of moral blame, including: (a) "not . . . tak[ing] pleasure in that person's successes, and not . . . hop[ing] that things go well for him;"[86] (b) withholding help with that person's projects (though Scanlon notes that it may remain inappropriate to withhold rescue or to go out of one's way to cause harm to those projects);[87] (c) "refus[ing] to make agreements with that person or enter into other specific relations that involve trust and reliance" and "suspen[ding] . . . friendly attitudes that signal a readiness" to enter into such relationships;[88] and (d) feeling "moral emotions or similar attitudes" such as "moral disapproval" and "resentment."[89] These suggestions are reasonable. Whether or not they are essential or central to blame, they are at least often associated with blaming.

However, Scanlon also believes that there are certain things we may *not* do when we blame. Scanlon rejects what he calls "moral retributivism," the

[86] Scanlon (2008) p.144.

[87] *Id.*

[88] *Id.* p.143.

[89] *Id.*

view that the proper response to blameworthy behavior is "to see even their most basic moral claims on the rest of us as limited and qualified."[90] He also rejects the idea that it is appropriate for a person "to suffer some loss in consequence" for blameworthy behavior.[91] Scanlon seems to see these two ideas as related. After stating that he rejects this idea—i.e., that it may be appropriate for a person to suffer a loss in response to blameworthy behavior—the reasons that he gives are all expressly directed at "moral retributivism."[92] The connection appears to be that believing it is appropriate for a person to suffer some loss also naturally involves believing that he has no claim against us for help in preventing that loss or for sympathy and concern about that loss and might even involve believing that he has no claim against our imposing or facilitating that loss.

When Scanlon argues against moral retributivism, he tends to present only an extreme instance of the view, one on which the loss to be suffered is death or serious harm. He claims that a view that supposes it is appropriate for a person to suffer such a loss would be implausible because we owe duties to help avoid such losses unconditionally: "it is implausible to hold that even the most basic moral requirements—such as the requirement not to inflict serious harm and to prevent such harm when one can—are conditional, and not owed to those whose attitudes impair their moral relations with others."[93] And again Scanlon states: "Even those who have no regard for the justifiability of their actions toward others retain their basic moral rights—they

---

[90] *Id.* p.142.

[91] *Id.* p.188.

[92] *Id.* p.189.

[93] *Id.*

still have claims on us not to be hurt or killed, to be helped when they are in dire need, and to have us honor promises we have made to them."[94]

But the view that Scanlon rejects on the basis of these arguments is much broader. The retributivism Scanlon rejects does not say that when we blame we may see it as appropriate for the person to suffer grievous harm or death and, therefore, as appropriate to withhold help in avoiding such harm. As Scanlon himself puts it, retributivism says more generally that when we blame we may see it as appropriate that a person should suffer some loss in consequence for his blameworthy act. One can accept this view without supposing that death or other forms of severe harm are ever among the appropriate losses. For one example, when we blame, we might feel that it would be appropriate for the person we blame to lose some sleep and some peace of mind over what he has done. There is nothing implausible about this. A person has no unconditional claim to our help in setting his mind at ease, nor even against being given a guilt trip. For another, when we blame a cheat or a scoundrel we might feel it is appropriate for him to suffer the loss of fortune and reputation in consequence for his actions. He also has no unconditional claim to our help in avoiding such losses. And for yet another, when we blame, we might feel that it is appropriate for a violent criminal to suffer the loss of his freedom. I see nothing implausible about the idea that one's right to freedom, and one's claims against others in this regard, are

[94] *Id.* p.142. The last idea thrown in here—that people who have no regard for reasons still have a claim against us to honor our promises to them—is far too sweeping to be credited. Suppose a woman promises to keep a man company for the day. She need not keep that promise if he becomes a boor and makes her feel unsafe. Or suppose that I gratuitously promise to give you five dollars if you will stop by my house to collect it, but when you do stop by, you kick my dog, set fire to the tree in my yard, smash up my car, and threaten death or worse for my family. The idea that you continue to have a claim on my five dollars is ludicrous. Promises are significant, but not that significant.

conditioned upon obedience to certain laws, including the avoidance of unjustified violence. So, even if we assume with Scanlon that people have (more or less) unconditional claims against death and severe harm, this does little to support his rejection of the more general view that it can be appropriate for a blameworthy person to suffer a loss in consequence for a wrongful act.

Scanlon also argues for his view by claiming that the blaming reactions he approves of can be justified by appeal to the notion of "appropriateness," whereas the "retributive" responses cannot:

> If blame involves only the alteration of attitudes that I have described, then it can be justified by appeal to the idea that this shift of attitudes is appropriate, or called for, by what the agent is like. . . . It is asking too much to demand that we be ready to enter into relations of trust and cooperation, and various forms of friendly relations, with people who have shown that they have no regard for our interests. Doing so can even be demeaning. So an appeal to what is appropriate is an adequate explanation for the suspension of these attitudes. But it is much less plausible to appeal simply to what is "appropriate" to justify the infliction of suffering on those who have treated others badly, or even to justify refusing to help them when they are in danger.[95]

Scanlon may be right that it would be asking too much to expect us to be ready to enter into relationships with those who have committed wrongful acts, but it is not clear why it would not also be "asking too much" that we respect their claims against all forms of suffering, while they cause us to suffer. Nor is it clear why it is not also "appropriate" to suppose that a person should suffer the loss of his freedom if he has used that freedom to harm others. Scanlon's

[95] Scanlon (2008) pp.189-90 (footnotes omitted).

appeals to appropriateness provide no answer to these sorts of questions because they are merely naked appeals to his own intuitions.[96]

Scanlon recognizes this as a potential problem but suggests that he can avoid it:

> Justifications that appeal to the idea of what is 'appropriate' or 'fitting' are open to the objection that they involve appeals to unstructured intuition, and unless supplemented in some way lack serious normative force. The view I am offering gives this idea more structure (thereby mitigating this objection, if not, to be sure, avoiding it altogether) by locating the idea of appropriateness with the conception of particular relationships, which explain the kind of normative force that is in question.[97]

But we have already seen that there is nothing to the "moral relationship" except the existence of moral norms and moral reasons. So, the idea of appropriateness here is "located," if anywhere, only within Scanlon's own views about the content of moral norms and moral reasons. That is to say, Scanlon's claims here are simply an appeal to his personal views about what we owe to each other unconditionally and what we do not.

On that front, Scanlon is out of step with common sense and with our actual attitudes and practices when we blame. We do commonly believe that it is appropriate, even just and good, for a blameworthy person to suffer, whether in the form of a guilty conscience or in the form of some more concrete loss. And we commonly believe that it is appropriate for blameworthy

---

[96] In *Moral Dimensions*, Scanlon notes that he previously stated his opposition to the view that it is sometimes the case that one should suffer in consequence for a blameworthy act in *What We Owe To Each Other* (Scanlon (1998)) at p.274. From a review of the section that Scanlon references I can discern no argument against that view, only a bald assertion that the view is "morally indefensible."

[97] Scanlon (2008) p.189.

people to be deprived of liberties, opportunities, and advantages. It is safe to say that the view that people do not deserve to suffer in consequence for blameworthy acts is far outside the mainstream. Our penal institutions and our public discourse about them make this quite clear.[98] But if you must, imagine a candidate for high political office proposing that rapists and murderers do not deserve to suffer in consequence for their actions. It would be political suicide.

Of course, this is not a conclusive argument that Scanlon is wrong. It is conceivable that these widely held views and common practices are wrong, but it would take more than Scanlon's assertions about appropriateness to show this. In any case, these claims are important because my immediate interest is not in deciding whether common sense views about blame are right or wrong, but in providing a reasonably accurate account of what they are. My goal is to provide a descriptive account of blame so that we can subsequently turn to the topic of whether this sort of blame can be defended against arguments based on determinism. Scanlon does not offer an account that captures our actual attitudes and practices when we blame.

IV. Blame and Blameworthiness: A Common Sense Retributive Account

In the foregoing, I have criticized a number of proposals for understanding blame. Along the way, I have made a number of claims about the nature of blame: that blaming involves more than a judgment identifying the person who deserves blame and more than a desire or wish that the person had not acted badly; that blame need not involve hostile feelings, like anger or resentment and that accounts that focus on reactive attitudes alone

[98] Our criminal law and penal institutions are publicly justified by and structured around the appropriateness of retributive punishment. To be sure, non-retributive justifications, such as deterrence and public safety, are mentioned and considered by legislatures, sentencing judges, and parole boards, but typically in conjunction with, not to the exclusion of, the virtually unquestioned premise that criminals deserve punishment.

(or on dispositions to have such attitudes) do not do justice to what I have called the normative significance of blame. I have indicated that I favor the idea, derived from P.F. Strawson and in some respects developed by Scanlon, that blame and blameworthiness involve a change to a person's moral standing, to the claims that he has and to the obligations to which he is subject. However, I have criticized Scanlon's attempt to develop this idea both because it imposes the unnecessary framework of a relationship on moral blame and because it dismisses, without reason, the retributive nature of blaming attitudes. In this section, I build on these claims and develop a descriptive account of blame and blameworthiness.

On my account, when a person is *blameworthy*, his moral standing changes in certain ways. By moral standing, I do not mean that his moral record declines, as though blame were a black mark on some account or grading sheet. Instead, I mean that he loses certain normative claims he otherwise would have had against others and that he is subject to new normative expectations and obligations to which he would not otherwise have been subject. For examples of lost claims, if a person is blameworthy for causing some injury to others, he will have less, if any, claim to the continued respect or friendship of those he has injured. He will also have little claim to sympathy or help in avoiding certain kinds of bad consequences that follow from his action, including consequences ranging from the resentment felt by his victims to the civil or criminal liability that might be imposed on him. Indeed, he may lose any claim he would have against our causing or facilitating these sorts of harsh consequences. In terms of additional expectations, we will also suppose that the blameworthy person owes his

victims an apology and that he may be obliged to take steps toward self-improvement in order to avoid committing similar acts.

Turning from blameworthiness to blame, on my account, to *blame* is to withdraw those attitudes or other things to which the person has lost his claim and to demand that he fulfill the special obligations that apply. So, the person who blames does not merely believe it would be appropriate to withdraw sympathy, for example, he actually withdraws at least some degree of sympathy. The person who blames also cares or insists that the blameworthy person meets the new obligations to which blameworthiness gives rise, such as obligations to apologize and to promise to avoid similar bad actions. The person who blames may insist that these obligations be met before he will accept that the blameworthy person should be permitted to return to his normal standing and before extending the sympathy that has been withdrawn.

With this rough statement of my view out of the way, let me try to be a bit more precise about some of the central changes that blameworthiness entails. Chief among them is the one that Scanlon rejects. When a person is blameworthy, it becomes appropriate that he should suffer in some degree in consequence. Thus, to some extent, he loses his claims to sympathy, to help in avoiding suffering, and to our avoidance of behavior that may cause him to suffer. As I argued in reply to Scanlon, this need not mean that it is ever appropriate for blameworthy person to suffer great harm. It need not even mean that it is appropriate for the blameworthy person to suffer any concrete loss or injury. In many cases, it is enough that the person should come to recognize the wrongness of what they have done and feel some remorse for it.

To be clear, shame, guilt, and remorse *are* ways of suffering for what one has done. These feelings do not always need to be excruciating or

enduring. If a person makes a rude or insensitive remark, she might "feel bad" about it without experiencing any intense suffering. That may be all the suffering she deserves to experience. On the other hand, we should not lose sight of the fact that guilt and shame can be terrible, destructive forces. In their extreme forms, they can eviscerate a person's sense of self-worth and lead him to do awful things to himself. I am not advocating the appropriateness of any particular degree of shame or remorse for any particular kind of blameworthy act. I simply wish to make clear that the idea that the blameworthy person deserves to feel some amount of remorse, shame, or guilt, an idea which I think is both plausible and commonly accepted, is not some weak impostor for the idea that the blameworthy person deserves to suffer.

At the same time, the view that a blameworthy person deserves to suffer should not be limited to suffering from guilt and shame. Sometimes, but certainly not always, we believe that a blameworthy person deserves to suffer in some more tangible or physical way. The belief that a criminal may deserve incarceration is perhaps the most obvious example, but it is not the only form that this attitude commonly takes. We commonly believe that, regardless of any state sanctioned punishment, a blameworthy person may deserve to lose his job, the love and trust of his spouse or his friends, his reputation, or his wealth, and we may hope that he does suffer some of these losses.

For an example, imagine a person who runs a business that he secretly knows to cause great and unjustifiable harm to others. Suppose that he has decided to have his company test dangerous chemical products or risky medical procedures on people in impoverished communities without their informed consent, while deceiving government regulators and corporate

shareholders about this practice. Seeing him as blameworthy, we may believe that he deserves to lose his job and any comparable position of trust and authority. We might also believe that it would be appropriate or well-deserved if he lost the wealth he has made from this business. These consequences seem quite fitting given that the blameworthy acts at issue arise in the exercise of his professional role. But we might also believe that he desrves other, less directly related consequences. We might judge it to be appropriate if his spouse, upon learning what he did, felt that he had so violated their shared principles that she could not trust or respect him again. We might see it as appropriate and deserved if she chose to leave him and raise their children without him. We may suppose not only that a blameworthy person brings these consequences upon himself, but that it is actually a *good* thing that he suffers them (or at least that it would be a bad thing if he did not).[99] As a result, by judging him to be blameworthy, we will suppose that he has no claim to help in avoiding these consequences and, indeed, no claim against reporting his conduct and thus causing or facilitating these harsh consequences.

[99] On Scanlon's view, it is inappropriate to see it as a good thing that a person should suffer a loss as a consequence for bad behavior. I find this view puzzling. Are we to suppose that it would be a good thing (or just an acceptable and fine thing) for the person in my example to go on happily without any consequence, to retain his job, enjoy his ill-gotten wealth, and continue to enjoy an unquestioning love and trust from family and friends? To avoid this implausible conclusion, Scanlon might appeal to intuitions that the impairment to the blameworthy person's relationship with his employer makes termination a fitting response, the impairment to his relationship with his wife makes separation or divorce appropriate, and the impairment to some relationship (perhaps with the employer or perhaps with his victims) makes it appropriate for him to lose his fortune (though note that I have not assumed that any deserved financial consequences must be brought about by repaying his victims; it may be appropriate enough that he loses his money to lawyers and creditors). But how does Scanlon explain how we, as third-party observers, can approve of these outcomes without undercutting his own rejection of retributivism? If we can see these as good or appropriate things, and surely we can, we should also be able to see it as "fitting" or "appropriate" for the public to impose a penalty in response to his violation of the public trust. But that is just to say we may see it as appropriate that people are made to suffer in certain ways as a consequence for blameworthy behavior.

The claim that a blameworthy person deserves to suffer and has no claims to sympathy or help in avoiding certain kinds of suffering can sound harsh and vindictive, but in fact, I suspect that the view could not be more commonplace. I, for one, am quite used to hearing, "Well, you should feel bad," as a response when I confide that I feel badly for having done something mean or thoughtless. The friends and family members who tell me this are not unusually cruel. It is quite normal to hope that people feel guilt and shame for their bad behavior. In fact, we often make a deliberate effort to provoke such feelings in people who have done something wrong. Though we might pejoratively call this a guilt trip, we very often seek to make a blameworthy person see the harm he has done from the perspective of others. Our goal is not simply to make him see it, but to make him see it *and* feel badly about it. After all, we will surely be disappointed and troubled if a new perspective on the pain does not cause remorse in the blameworthy person, but instead leaves him indifferent or even more inclined to do it again. This appears to indicate that we believe that blameworthy people ought to suffer remorse, at least, and that they have lost claims against being made to feel remorse.

Of course there are limits to the ways in which blameworthiness should affect our view of a person's claims. Retributivism should not be mistaken for the view that punishment and suffering are appropriate without qualification. How much a blameworthy person deserves to suffer depends on what she has done. So, though we may believe that a person deserves to suffer to some degree and though she may have no claim to our sympathy if she suffers in the appropriate degree, she may deserve every bit of our sympathy and assistance if her suffering becomes disproportionate. If she is *made* to suffer disproportionately, this may be cause for disapproval, blame, and even

outrage. We may believe that the immoral businessman in my example deserves to suffer in all of the very serious ways described above, and yet disapprove if he receives a punishment that offers little opportunity for redemption, and we may be horrified if he is attacked and brutalized by an angry mob. In short, whatever claims a blameworthy person loses, he or she certainly retains claims against excessive punishment.

In addition to believing that it is appropriate for a blameworthy person to suffer, the blameworthy person may lose a number of other claims that he may normally have on us. As indicated above, Scanlon identifies a number of these sorts of claims. There may be others. To catalog them with particularity would be difficult, as they may vary greatly with the circumstances and nature of the blameworthy act. But in general terms, a blameworthy person *may* lose his claims to our trust, our esteem, our friendly or cordial attitudes, and our openness and willingness to engage and work with him or to enter into relationships. Where he does, this is not to say that we must withdraw our trust, esteem, etc. For one, the point is that the blameworthy person loses his claim to these things, not that we have no right to extend them if we choose. But in addition, in viewing a person as blameworthy, we may believe that the changes to the person's claims and obligations should be localized in particular ways. In judging one's boss to be blameworthy for infidelity to her husband, we may feel that the boss has lost her claim to *her husband's* trust and to the respect and friendliness of *his* family and friends. So, we may believe it is appropriate for *them* to withdraw from her in these ways. But we may still feel that our boss is entitled to respect from people in the workplace and to trust and esteem in matters involving the business.

As I have said above, a blameworthy person does not only lose claims that she might otherwise have. She also becomes subject to special normative expectations or demands. Typically, blameworthiness includes or gives rise to obligations to apologize or otherwise show remorse and contrition; obligations to make reparations and correct or offset, to some reasonable extent, the effects of the bad behavior; and obligations to take measures to avoid similar behavior, whether that involves merely making a resolution or something more, like habit formation or treatment. When judging a person to be blameworthy, we accept that she is appropriately subject to some or all of these obligations. However, in simply judging her to be blameworthy, we need not care particularly that these obligations are met. We may feel it is none of our business. But we will believe that she deserves to be held to these expectations by appropriate blamers. The difference between judging a person to be blameworthy and actually blaming him is, on my view, caring in the particular instance that the person gets what he now deserves and that he fulfills these new obligations. When we blame, we do not merely judge that it would be appropriate for the blameworthy person to suffer in some degree, we actually want him to suffer. We expect the person to feel badly for what he did, whether this means only a moment of understanding and remorse or years of guilt. It may cause us pain to see the blameworthy person continuing to thrive in ways that we feel he no longer deserves. And to the extent that the person has lost his claim to our trust, esteem, etc., then blaming involves actually withdrawing our trust, esteem, etc. Blaming involves imposing the sorts of special obligations identified above, and actually caring that they are fulfilled. When we blame, we do not merely suppose that it would be appropriate for the person to apologize and fix the damage he has done.

We care about whether or not the person does so. We may refuse to extend our trust, sympathy or openness to him until he meets these expectations, and if he does not, we may take offense and sharpen our blaming responses.

The various special expectations associated with blame and the belief that the blameworthy person deserves to suffer are not only part of what it is to blame, they are conditions for the termination of blame and blameworthiness. If a person has adequately fulfilled these expectations—if say, he has suffered remorse, shown sufficient contrition, and made an honest effort to better himself in ways that will prevent similar blameworthy acts—it typically becomes appropriate to stop blaming. The person who has suffered enough and who has improved or reinvented himself may have a claim on others to stop blaming him. If not in every case, then in all but those involving egregious behavior, it will be possible for these special expectations to be met during the blameworthy person's lifetime. In many cases, the appropriate demands can be met very quickly and easily, as they might require only a short period of remorse, an apology, and a promise not to repeat the bad act. Even in more serious cases, it will normally be possible to reach a point where one no longer deserves blame for some instance of past behavior. Thus, blameworthiness is often, if not always, a temporary and terminable change to one's moral standing.

The foregoing description of blameworthiness and blame supports the fact that when we blame we very often feel certain reactive attitudes. It is not hard to see that behavior that would make us want a person to feel and show remorse would also naturally cause anger and indignation, or that the belief that we should withdraw trust from a person whom we had previously liked could be accompanied by disappointment and sadness. And if blame involves

a withdrawal of sympathy and concern, it naturally will also lead to outright antipathy in many cases. But my view of blame does not make these feelings a prerequisite for blame. I believe that attention to the normative implications of blame, to the ways in which we alter our view of the claims and expectations that apply, shows that these feelings, however important they are, are not the essence of what it is to blame a person.

My account of blame also naturally makes room for the practice of retributive punishment. Retributive punishment can be justified only if it is appropriate that a person suffers in consequence for blameworthy behavior. That alone is not enough, of course. For one thing, we could think it appropriate for a person to suffer for what she's done without thinking that anyone has the right to make her suffer. Our aversions to hypocrisy, vigilantism, and penal systems that lack procedural safeguards, all show that we believe that not just anyone should be empowered to punish. In addition, in many cases at least, we may believe that a person has lost her claim to sympathy or to help in avoiding suffering without also believing she deserves punishment of any of the sorts normally imposed by the state. My conception of blame supports the propriety of retributive punishment, but it is not sufficient for it. In other words, on my account, in any particular case, it does not follow from the fact that a person is blameworthy that she should be punished, and a further account of the conditions for the justifiability of retributive punishment would need to be developed in order to explain whether retributive punishment is ever appropriate. My retributive account of blame and blameworthiness merely allows room for the possibility of that retributive punishment is appropriate.

Having discussed blame and blameworthiness, it is worth adding that there is room for an analogous account of praiseworthiness. More so than "blame," the term "praise" calls to mind a communicative act, speaking well of someone. "Praise" may not commonly be used in a way that corresponds to blame in the robust, normative sense that I have identified here. But even if "praise" refers most naturally to a speech act, praiseworthiness can be understood as involving a change in moral standing, analogous to blameworthiness. To deem a person to be morally praiseworthy is not simply to say that she deserves to be praised outwardly, in the speech act sense. It involves believing that she deserves to be held in esteem, to be trusted and granted a degree of deference in certain matters, and perhaps to enjoy our gratitude. Praiseworthy people deserve particularly strong concern and sympathy. It seems particularly sad or repugnant when they suffer bad luck or when they are treated without respect. We feel that in a just world they should not suffer, but instead be rewarded for their virtue. Thus, when one is praiseworthy, one has some claim to heightened respect, trust, and concern from others. Of course, praiseworthiness, like blameworthiness, is a terminable condition. One deserves only so much credit for individual praiseworthy acts, and if one ceases to behave admirably, one will eventually cease to be praiseworthy.

V. The Possibility of Real Moral Responsibility

I began this Chapter with the claim that an understanding of moral responsibility is crucial to evaluating arguments about the conditions for moral responsibility. To demonstrate that fact I began with a look at Galen Strawson's argument for the impossibility of "true" moral responsibility. There, I relied solely on intuitions about particular cases of blameworthiness. Now, I

want to return to Strawson's argument and explain why it does not apply to blameworthiness and praiseworthiness of the sort that I have now identified. Recall that Galen Strawson's argument depends on identifying certain conditions for responsibility, particularly the supposed condition that a person must have chosen, consciously and explicitly, the mental states that caused or motivated the behavior at issue. To respond to this suggestion, this section explains what my account of blameworthiness and praiseworthiness indicates about the conditions for moral responsibility. That is, it offers an initial answer to the question, "What must be true of our behavior for it to warrant the blaming and praising responses that I have just described?"

My account of praise and blame supports a view of the following sort: moral praise and blame for behavior may be deserved only as a response to the strength of that person's moral motivation evident in her behavior. A person deserves blame for her behavior, *i.e.*, is blameworthy, only if that behavior reveals that her concern for conforming her behavior to normative reasons is weak, and a person deserves praise for her behavior, *i.e.*, is praiseworthy, only if that behavior reveals a particularly strong concern for normative reasons. By strong and weak, I do not mean strong and weak in relation to that person's other motives at the time. One does not deserve much praise for doing the right thing simply when all of the other options are not very attractive, or where no normal person would be tempted to do otherwise, and one does not deserve blame if one is deeply concerned to do the right thing, but that concern is overwhelmed by a truly pathological desire. Instead, strength and weakness must be measured relative to that degree of moral motivation which we may reasonably expect of a person. If a person's actions reveal that she is more motivated to do the right thing than we would

reasonably expect, then she may be praiseworthy, and if they reveal that she is less motivated than we should reasonably expect, then she may be blameworthy.

On this view, moral motivation of the sort that is sufficient to avoid blameworthiness neither entails, nor is entailed by what I referred to in Chapter One as "sufficient motivation" to do the right thing. "Sufficient motivation," as I used the phrase there, means motivation that is actually sufficient to trigger an exercise of a given ability. But as I have just indicated, a person may lack sufficient motivation in that sense, perhaps because of an opposing pathological desire, and yet not be blameworthy. When we are concerned with what a person ought to do, we ask what they would have done if they were sufficiently motivated to do the right thing. That is, we consider what they would have done if their motivation to act on reasons outstripped their other desires and temptations no matter how strong those other desires are. So, it will be the case that a person ought to have done something else any time that his moral motivation is weaker than a desire that impels him to do what he does. But we will judge him to be blameworthy only to the extent that his moral motivation falls short of what we can reasonably expect. Thus, a kleptomaniac ought not steal, but she may not deserve blame if she makes a heroic, but unsuccessful effort to avoid stealing. We might even see that as an occasion for praise.

Similarly, a person who has very strong, but not pathological, motivating reasons for doing the wrong thing could be less blameworthy than the person who does the wrong thing without any need to do so. Compare a person who trespasses into a building because he desperately needs a bathroom with a person who does so simply because she wants to look around. The first

person might be blameworthy. Perhaps he could have planned better and avoided that situation. But the second person is probably more blameworthy. We may infer that the second person's desire to act on the relevant normative reasons is very weak. From the fact that her desire to look around took precedence, we may infer that she is utterly unconcerned about the normative reasons not to trespass onto another person's property. But we cannot infer the same about the first person. We only know that his respect for the reasons not to trespass gives way in cases of somewhat more serious need.

Let me elaborate on my proposal a bit more and then explain why I take it to follow from the account of moral responsibility that I have developed here. We may reasonably expect moral agents to have some general moral motivation, some general desire to do what they ought to do. This might be phrased in a number of ways: a desire to do the right thing, a desire to do what one has most reason to do, a desire to pursue value, a desire to be good, etc. For my purposes here, I am not aware that there is any significant difference between these descriptions of a basic moral desire. The inculcation or encouragement of this general desire, and the suppression of its opposite, is a central part of any minimally adequate moral education.[100] Setting worries about determinism to the side for a moment, it is presumably reasonable to

[100] Of course, we may also reasonably expect moral agents to have other more specific, but still quite general moral desires and dispositions, such as desires to be fair, kind, courageous, just, honest, etc. But I am going to focus on the more general desire. It is difficult to identify a virtue of this sort that cannot be overdone and thus lead to non-optimal behavior. People can be too kind, too courageous, etc., unless these desires are moderated by a more general interest in doing the right thing. So, simply acting from a desire to be kind may not be acting from a desire to do the right thing, nor something that merits praise. But I do not mean to exclude these dispositions from the idea of moral motivation. I assume that any reasonably adequate concern to do the right thing will naturally provide a motive for developing these more specific desires and dispositions (to the right degree) and suppressing their opposites. So acting as a result of these dispositions can be a way of revealing the strength of one's concern to do the right thing.

expect any moral agent to be strongly but, given other natural human desires and temptations, imperfectly motivated by this moral desire. On my view, only behavior that reveals that an agent is weakly motivated by this general moral desire is behavior for which she may be blameworthy.

However, my proposal is not that we should assess blameworthiness and praiseworthiness simply based on the strength of a person's moral motivation at the moment before acting. Very often it is not obvious what the right thing is, and in assessing responsibility, we give people the benefit of some, but not all, of their mistaken beliefs about what it is. As a result, in some cases, a person may be blameworthy for behavior even though they were, in that instance, motivated by the desire to do the right thing. This does not conflict with my proposal. If it is reasonable to expect a person to be motivated by the desire to do the right thing, it is equally reasonable to expect him to want to find out what the right thing is. In fact, one cannot fully have the former desire without the latter. As a result, a person has a duty of reasonable care to investigate and think through his beliefs, both factual and moral, and to discern the truth about what he ought to do.[101] The diligence with which a person performs this duty may then affect his culpability for his behavior in other instances. His later behavior may reveal or support an inference that he has not been concerned to think about and figure out what he ought to do.

There is another way in which one's moral motivation at the instant of acting is not a sufficient guide to blameworthiness. Knowing what one ought

[101] The point of calling it a duty of reasonable care is that the scope of the duty is limited not only by the agent's ability and opportunity to ascertain the truth, but also by a reasonable balancing of the foreseeable importance of a particular area of facts as compared to the cost of ascertaining the truth about it. So, though an agent might have the ability and opportunity to investigate the potential distant consequences of his behavior, he has no duty to do so where the information is likely to be of very little practical value and gathering it will interfere with other more important activities.

to do is not always a matter of puzzling through tough moral questions and balancing competing reasons. Sometimes it just requires remembering one's promises and planning so that one can fulfill them. If a person makes a commitment, but then fails to take reasonable measures to ensure that he will remember that commitment, he will be blameworthy if he forgets to fulfill that commitment. This is so even though when the time comes to fulfill the commitment his intentions are perfectly good. So, like the person who neglects to reflect on his moral views, this person may be blameworthy even while he is currently motivated by a desire to do the right thing. And as in that case, this is reconciled with my view by the fact that strong motivation to do the right thing includes the desire to ensure that we are later in a position to do the right thing. Our failure to remember a promise may reveal our failure to plan ahead or to set ourselves a reminder, and that in turn may reveal a lack of concern to do the right thing.

What these points show is that we cannot evaluate the strength of a person's concern to do the right thing by focusing narrowly on the immediate source of his motivation. In looking at what a person's behavior reveals about the strength of his concern to do the right thing, we take a broader view. We consider what their behavior shows about their desire to figure out what they ought to do and to put themselves in a position to do it. On my view, we cannot make an accurate judgment about how concerned they are about doing the right thing unless we take this broader view.

My proposal is similar in spirit to the Kantian position advocated by Nomy Arpaly, but different in detail. Arpaly suggests that "[a] person is praiseworthy for taking a morally right course of action out of good will and blameworthy for taking a morally wrong course of action out of lack of good

will or out of ill will."[102] This account correctly directs our attention to the quality of the person's motives and desires, but it has two disadvantages. As I have just argued, a person can be blameworthy despite being, at the moment, sincerely motivated by the desire to do the right thing and thus apparently acting with "good will." Perhaps she has previously been lazy about figuring out what the right thing is, or perhaps she has failed to be vigilant about remembering her obligations. There may be wiggle room in the idea of taking a course of action "out of good will" that would allow Arpaly to account for this, but that would only emphasize the second problem, the vagueness of these terms. Though I suspect I have some sense of what is meant by "good will" and "ill will," I do not see their vagueness as a virtue. In place of these terms, my view specifies that in assessing blameworthiness and praiseworthiness for actions, we are concerned with what those actions reveal about the strength of a person's motivation to do the right thing, in the broad sense that includes her desire to discover reasons and determine how they apply in a given situation and to plan and prepare so that she can be in a position to do what she should, and that we are concerned with the strength of this desire relative to our reasonable expectations, rather than relative to her other desires.

The view that blameworthiness and praiseworthiness depend on what the agent's behavior reveals about the strength of her concern for doing the right thing gains support from my account of blame and praise. Moral praise and blame involve departures from the attitudes that we normally ought to have towards one another. They are departures from the respect, sympathy, and good will that people generally deserve. On this account, a person's claims to respect, sympathy and concern are to a certain extent conditional.

[102] Arpaly (2006) p.15.

They may lose those claims—or in cases of praise they may earn a stronger claim—based on what they do. Moral blame also involves the imposition of additional obligations or expectations upon the blameworthy person. This raises the following questions: On what grounds would it make sense to suppose that a person could lose or gain such claims, and on what grounds may we make these additional moral demands of people? On what grounds would it make sense to say that a person's behavior makes them deserve a departure from these basic moral attitudes? I believe the answer is only on the grounds that their behavior shows that their concern to do the right thing has exceeded or fallen short of that which can be reasonably expected from them.

Consider moral blameworthiness first. We normally owe each other a reasonable degree of sympathy and concern, but on the view that I have argued for here blame involves believing that a person should suffer. Moral blame involves a change to attitudes that people normally deserve solely in virtue of their standing as fellow persons. It is difficult to see how someone could deserve a departure from these attitudes unless his behavior reveals that he does not care very much for moral reasons himself. It is a basic moral intuition, I think, that a person would not deserve a withdrawal of moral respect and concern and may not be held to heightened moral obligations simply because she is weak, slow, unattractive, or unlucky, or because she has any other morally insignificant trait. Instead, her normal moral standing depends only her status as a person and on the presumption that she is reasonably concerned for moral and normative reasons. Only if her behavior displays a failure to be concerned for and motivated by moral reasons and moral value—

and not simply a lack of intelligence, skill, or resources—should we say that she deserves any alteration to her normal moral standing.

This might sound a bit like a crude eye-for-an-eye retributive principle: if she does not show others moral respect, then we need not show it to her. My account of blame and blameworthiness is intended to be retributive, but not crudely so. Any impression that I am relying on eye-for-eye type reasoning overlooks important differences between blameworthy behavior and blaming responses. Blameworthiness affects *some* moral claims and supports a withdrawal of a *degree* of moral concern and respect, but blameworthiness and blame are not simple matters of settling the score. As I claimed above, even very blameworthy people still have claims against many forms of cruel treatment. The idea that a blameworthy person may deserve to suffer in some ways does not mean that he deserves anything like the behavior for which he is blamed. Thus, a torturer may deserve to suffer, but he may maintain a claim against being tortured.

The conditions that call for an end to blame also confirm both that blameworthiness is not crudely retributive and that blameworthiness depends on the quality of a person's motives. Apology, remorse, and demonstrated change of character can make it reasonable to decide that a person no longer deserves blame. They do so because they reveal that the formerly blameworthy person has come to appreciate the problems with her prior behavior and would no longer want to behave that way if she had another chance. Her remorse and her efforts to change show her concern for reasons. Moreover, in such cases, it is often appropriate to stop blaming regardless of whether or not the "score has been settled" and our vindictive impulses have been satisfied.

It must be admitted that there are cases in which it is natural and understandable to blame a person although her behavior cannot be explained by a lack of concern for reasons. We sometimes feel bitterness and resentment towards a person who has harmed our interests, even if the harm was just the result of fair competition or bad luck. Their close causal connection with our loss may make it difficult to maintain the respect and concern we ought to have towards them. But the fact that this is natural or understandable does not mean that the person *deserves* anything less than full moral respect. It simply means that we can see why we might be prone to respond in this way.

Similarly, there are cases in which our blaming attitudes are in part responses to a moral fault, but also are affected by other factors, such as luck. We tend to blame the drunk driver who injures a person more than the equally drunk driver who, luckily, makes it home without causing harm. My own view is that these responses, like those above, are *understandable*, but that they do not correspond to what is *deserved*. We have understandable tendencies both to be harsh, possibly too harsh, to the unlucky drunk driver and especially to be too easy on the lucky one. Blaming can be emotionally and practically difficult, particularly where it would involve distancing ourselves from someone that we care about or with whom we cannot avoid interacting. So, we may be prone to let a person off the hook easily where he has done no real harm. But it is more difficult to avoid withdrawing from and feeling resentment towards a person who is closely, causally linked to a serious and vivid moral harm.

Now consider praiseworthiness. The appropriateness of some of the attitudes that a praiseworthy person deserves obviously depend upon the strength of her concern to do the right thing. The esteem she may deserve is

not esteem for her athletic prowess, artistic ability, or mental acuity, but specifically for her moral character and her interest and resolve to do the right thing. Similarly, the trust and deference a praiseworthy person deserves relates to the fact that she has shown herself to have good moral character, not some technical skill. We should trust her to try to discern her duty and carry it out, not to score points in a basketball game or bake a soufflé. Such trust could only be deserved on the basis of behavior that shows she is strongly concerned about acting on good reasons. Our belief that a praiseworthy person deserves happiness also depends, though less obviously so, on the fact that their action displays strong concern for reasons. We generally believe that people of skill, talent and natural ability may deserve to be rewarded. They deserve to be fairly compensated for their productive talents, which is to say that they deserve reward in return for the value and happiness that these talents bring to others. But this kind of desert is different than the desert of happiness that comes with moral praiseworthiness. What people deserve in the first instance depends on the existence of a market for the products of their skills and talents, and it is desert of compensation, not happiness. Unlike a person of skill, a person of moral character and strong concern for reasons deserves happiness, and she deserves it regardless of whether her virtue is appreciated or economically valued by others. This could only be because her behavior shows exceptional concern to do the right thing.

Now let us consider Galen Strawson's position that to be responsible for our behavior, we must have chosen the mental states that caused that behavior. As I have just argued, from the nature of blame and praise, we can see that judging a person to be blameworthy or praiseworthy for her behavior requires us to be able to make a judgment about the strength of her concern

for reasons relative to our reasonable expectations. But this conception of praise and blame does not give us any reason to believe that blameworthiness or praiseworthiness requires us to know anything more about the mental causes of a person's behavior. For unlike heaven and hell, praise and blame are generally temporary, terminable responses to a person's behavior. We need not make any *final* assessment about a person's moral character in order to judge her to be blameworthy or praiseworthy. Judgments that a person is blameworthy or praiseworthy may not be fleeting, but they are subject to revision as the agent reacts to her own behavior and as she continues to act in ways that further reveal her character. We can also blame or praise for behavior without making any *overall* assessment of a person's moral character. People who are generally good people can deserve blame for isolated instances of bad behavior. And people may deserve blame for a lack of moral motivation that is evident in one field (*e.g.*, laziness in work and career), while simultaneously deserving praise for showing strong moral motivation in another (*e.g.*, kindness and sensitivity towards others). Thus, there is nothing inconsistent about thinking that a person deserves some degree of blame, while thinking she is, by and large, a decent or good person. Nor is there any inconsistency in believing that a person should feel remorse and that she should learn from her mistakes, while also believing that she is the sort of person who will do so quickly. Thus, to support praise or blame, we need not believe that a person's behavior accurately reflects something deep about who she truly or fundamentally is. We need only determine that she ought to have been more concerned to act on reasons, or in the case of praise, that she was more concerned than we might have expected.

This is not to say that having or lacking a certain degree of moral concern is sufficient for praiseworthiness or blameworthiness. Perhaps we can argue that other conditions pertain. Nor should I be taken to be suggesting, at this point, that determinism does not pose a problem for moral responsibility. My explanation of when a person deserves blame depends on the idea that we could reasonably expect a person's moral motivation to be stronger. One might ask when it is reasonable to have that expectation, particularly in a deterministic world. I have passed over this issue so far. My goal here is not to identify all of the conditions for responsibility. It is to explain why Galen Strawson's conditions do not apply. My conception of blame and praise provide no reason to suppose that real moral responsibility, including praise and blame, requires that a person consciously and explicitly *chose* the degree of moral motivation that they have, as Galen Strawson assumes. Having developed an account of moral responsibility in terms of blameworthiness and praiseworthiness, we may now turn to some of the other arguments that suggest determinism might undermine moral responsibility.

CHAPTER THREE

MORAL RESPONSIBILITY AND THE POSSIBILITY OF BEHAVING DIFFERENTLY

I. Introduction:

A person is not blameworthy for what she does if it would be unreasonable to expect anything better of her. Similarly, we might suppose that a person has no claim to praise or admiration if no one in her situation would have been tempted to do less. Taken together, a person is morally responsible for what she does only when it is reasonable to expect something different from her. These claims, I am confident, are true when properly understood. However, they come close to implying a familiar and highly controversial claim about responsibility. If we can reasonably expect better (or worse) from a person than what she has actually done, then we are only a small step from supposing that she could have *done* something else. Thus, we are naturally led to the principle of alternate possibilities (PAP): If a person is morally responsible for what she has done, then she could have done otherwise.

Stated this way, PAP is weaker in one respect than the intuitions from which I started. It does not specifically require that the possible alternative is better for cases of blameworthiness and worse for cases of praiseworthiness, and it does not require that the alternatives are within the range of our reasonable expectations. It requires only that some alternative course of

action is possible. Nonetheless, PAP is strong enough to pose a problem for anyone who supposes that morally responsibility and determinism are compatible. PAP provides the key premise in an argument much like the one we confronted in Chapter One. That argument might be put this way:

1) If determinism is true, then no one could have done anything other than what he did.

2) If a person is morally responsible for what she has done, then she could have done something else.

3) So, if determinism is true, no one is ever morally responsible.

Based on the discussion of the argument from determinism in Chapter One, it is clear that more needs to be said about the sense of "could have" in premises 1 and 2 before we declare this argument to be valid. Still, the intuitive plausibility of the premises and the appearance of validity are enough to make clear that any defender of the compatibility of determinism and moral responsibility must have something to say about the principle of alternate possibilities.

There are two widely recognized compatibilist responses. The first—the "traditional compatibilist" response—accepts the truth of PAP, but exploits the flexibility in the meaning of "can" or "could have" in order to block the argument. The most familiar versions of this strategy suggest that, as used in the principle, "could have done otherwise" means only "would have done

otherwise, if she so chose."[103] If some such interpretation of the principle can be defended, then the argument is not valid. As we have already seen, determinism implies that no one could have done otherwise in a different and stronger sense than this.

The second compatibilist strategy does not attempt to finesse the principle in this way. It rejects the principle outright. Those who reject PAP generally do so on the basis of a controversial counterexample developed by Henry Frankfurt and elaborated by countless critics and supporters.[104] This may be called the "Frankfurtian compatibilist" response.

To generate a Frankfurt-type counterexample, we start with a person who is morally responsible for what she does. Suppose, for example, that a competent moral agent, J, does something awful and that she does it for utterly bad reasons. We then simply add that, unbeknownst to J, there was someone else, B, who was ready to intervene if necessary to make sure that J "does" what she actually did. So, B would have forced J's hand if J had changed her mind or otherwise became inclined to do the right thing. But because J chooses to act for her own bad reasons, B does not actually need to intervene. In such cases, Frankfurt and his supporters suggest that we should see J as blameworthy because she acted for her own bad reasons. The fact that B was ready to intervene had nothing to do with what she

[103] Conditional analyses of this sort have been offered for various principles related to freedom and moral responsibility including the principle of alternate possibilities. *See, e.g.* Hume (1740) p.63; Moore (1912) pp.13, 196-222; Austin (1956); Ayer (1954).

[104] *See* Frankfurt (1969).

actually did, and therefore nothing to do with her responsibility for what she did. However, they also insist that B's possible intervention ensures that J could not have done anything else. The key to these cases is the insight that something can make an agent unable to do anything but φ without itself causing her to do φ.

In the literature on Frankfurt cases, there has been a great deal of disagreement about whether they really rule out the possibility of doing otherwise. Everyone must concede that there is inevitably *some* difference between the course of events in which J acts for her own reasons and the alternative in which B intervenes. In the counterfactual scenario in which B intervenes, J's behavior has different causes; J's reasons for acting in the actual scenario do not motivate or explain her counterfactual behavior; and J is almost certainly not responsible for what B makes her do. Frankfurt's critics sometimes stress the importance of these differences, arguing that they show that J has alternatives. Frankfurt's supporters, on the other hand, have downplayed the significance of these differences and insisted that they do not amount to doing otherwise in any sense that could be relevant to moral responsibility.

There is something to be said for each side. It should be conceded that the differences between acting on one's own and being manipulated by a Frankfurtian intervener are not sufficient to sustain the principle of alternate possibilities as stated above. This is not because Frankfurtian interveners are able to ensure that the agent does the same thing in the counterfactual

scenario, but because, when they intervene, they can ensure that there is nothing whatsoever that the agent *does*. If there is nothing that an agent does, then she is not *doing* otherwise. Frankfurt's supporters go wrong, however, if they suppose that this means the alternative, counterfactual scenario is irrelevant or unimportant to the agent's responsibility for her behavior. Alternate possibilities of a sort do remain, and they are essential to moral responsibility. The fact that Frankfurt cases, no matter how they are tweaked, cannot rule out differences between the agent's actual and counterfactual behavior should be a clue that these differences reflect an important truth about moral responsibility. Once these differences are accurately characterized and understood, rather than downplayed or dismissed, it should be clear that a principle somewhat like PAP states a necessary condition for moral responsibility. I refer to this principle as the Principle of Alternate Possible Behavior (PAPB), and I will argue that we have reason to accept a traditional compatibilist version of this alternative principle.

II. The Principle of Alternate Possible Behavior

A. Introduction to the Principle of Alternate Possible Behavior:

PAP says that if a person is morally responsible for her behavior, then she could have done otherwise. It makes the possibility of doing otherwise a necessary condition for responsibility for actions and omissions. This section explains and then defends a different principle, PAPB. Here is a rough statement of PAPB, which will need some explanation and refinement:

> If a person is morally responsible for her behavior,
> then she could have behaved differently.

The key difference between PAP and PAPB is that PAPB does not require a counterfactual scenario in which the agent *does* anything at all. Instead, it only requires an alternate possible scenario in which the agent's behavior is relevantly different. To bring out this difference between PAP and PAPB, let us define some terminology and then consider an illustration.

PAP relies on the possibility of *doing* otherwise. I understand the things that an agent *does* to include her voluntary actions and her voluntary omissions, but nothing else. Including omissions within the category of things a person does may seem strange or even like an oxymoron. For some purposes it would be a mistake not to distinguish between actions and omissions, but for my purposes here I see no serious problem with lumping them together.[105] An omission may be thought of as a case where one allows something to happen or allows something to be the case though one has some opportunity to prevent or change it. Allowing is something that a person "does" in the sense that I am concerned with.

Actions and omissions are *voluntary* in the weak sense that I have in mind if they can be explained by an agent's beliefs, desires and other

[105] Note the significance of calling a voluntary omission something that the agent does. For one thing, it means that I read PAP as stating a single condition on responsibility for both actions and omissions. Not everyone has supposed that this is so. Van Inwagen, for example, has suggested that PAP concerns only "performed acts," not omissions. Van Inwagen (1983) pp.164-65. Incidentally, Van Inwagen is willing to concede PAP to the Frankfurt examples, but not his "Principle of Possible Action," which states that a "person is morally responsible for failing to perform a given act only if he could have performed that act." On my approach, the Principle of Possible Action is simply a more restricted version of PAP.

motivational states in the normal, familiar manner. This means that an act or omission may be voluntary although it is done only in response to a threat or compulsory legal order. A person who, fearing a death threat, does something that she finds loathsome still does that thing voluntarily on this account. Actions and omissions may also be voluntary even if they are brought about only through a mistaken belief. The person who shoots a friend, mistaking him for an enemy, does so voluntarily even though he may rightly point out that he did not mean to do it.[106] It should also be noted that an omission may be voluntary not only if it is affirmatively motivated by the agent's desires (*e.g.*, the omission occurs because the agent is motivated to do something else), but also if it can be explained by the absence or weakness of her motivation to perform the relevant act (*e.g.*, the omission occurs because, lacking concern about her duty, the agent lies around and does nothing at all).

Unlike PAP, PAPB draws our attention to *behavior* and to the possibility of *behaving differently*. I understand "behavior" to cover far more than voluntary actions and omissions. "Behavior" also includes any non-voluntary motions of an agent's body, as well as the absence thereof. Twitches, snores, and movements made in sleep are a few familiar examples of behavior. Behavior also includes falling over as the result of a push, simply remaining motionless, and the rising and falling of one's chest from breathing. The fact that PAPB is addressed not only to what an agent does, but also to the rest of

[106] Thus, "voluntarily" is weaker than "intentionally." I suspect that an action will be voluntary under each of its descriptions if it is voluntary under any of them, whereas an action can be intentional under only those descriptions that the agent believes to apply.

her behavior does not mean that I suppose that a person is likely to be responsible for non-voluntary behavior. Indeed, PAPB itself indicates that typically, if not always, non-voluntary behavior is not the sort of behavior for which we could be blameworthy.

Now, to illustrate the difference between PAP and PAPB, imagine that Jane faces a choice. She must decide whether to submit her dissertation tomorrow or to delay, perhaps indefinitely (*i.e.*, not submit tomorrow). (Submitting today is not an option because Jane is in the wilderness on a camping trip.) Normally, Jane will do one or the other: submit or delay. But it is also possible that before she does either of these things, Jane falls into a coma that lasts for weeks. If so, then submitting her dissertation will not be among the things she does tomorrow, but it also will not be the case that she *does* otherwise. There is nothing that she does once she falls into the coma.

Of course, it would be correct to say that she does not submit her dissertation, but not if this is interpreted to mean that she *does* something other than submit her dissertation. There is an implicit "voluntarily" here, and we must be careful about where it goes. It is correct to say that Jane does not voluntarily submit her dissertation, but it would be wrong to say that Jane voluntarily does not submit her dissertation.

So, while Jane is on her camping trip, still healthy and conscious, at least three courses of events seem possible: she might voluntarily submit, she might voluntarily not submit, or there might be nothing that she does voluntarily. The first two confront her as choices. The third does not. It arises

in a case where her agency is suspended, so to speak. These three possibilities bring out the difference between PAP and PAPB. PAP focuses our attention on only the first two options. It suggests that if Jane is praiseworthy for submitting, then it must have been possible for her to delay (*i.e.*, voluntarily not submit). PAPB, on the other hand, asks us to consider all three possibilities. According to PAPB, Jane may be praiseworthy as long as there is a possible scenario in which she fails to submit, whether this is because she voluntarily delays or because there is nothing she does at all.

The claim that there is a third possibility not treated as relevant by the normal principle of alternate possibilities might seem surprising. One might suggest that, in PAP, "could have done otherwise" is intended to cover the full range of possibilities, including the possibility of doing nothing voluntarily whatsoever. After all, there is a loose sense in which a person's non-voluntary behavior can be included among the things she "does." We might ask a person, "Do you realize what you did in your sleep last night?" (Of course, we can talk this way about inanimate objects as well: "Watch what this thing does.") If one wishes to use "does" in this sense, I could concede that PAPB is merely a restatement of PAP that helps to bring out this third possibility. But I do not believe that this is how "could have done otherwise" has normally been understood or that it is faithful to the supporters of PAP. This is most clear on the traditional compatibilist interpretation of PAP. Traditional compatibilists claim that "could have done otherwise" means something like "would have done otherwise, if she had so chosen." This interpretation makes

explicit the restriction of relevant alternatives to those in which the alternative is one that may be chosen, or in other words, to those in which the alternative involves voluntary behavior. This restriction is often made explicit by those who reject the traditional compatibilist interpretation as well. The customary rejoinder is that "could have done otherwise" means not only "would have done otherwise if she had so chosen" but in addition ". . . and nothing prevented her from so choosing." Furthermore, as I will argue below, if people had understood PAP to be satisfied by the possibility of not doing anything voluntarily, they should not have supposed that Frankfurt's case presents a counterexample. Therefore, I understand PAPB to be distinct from PAP.

Two further refinements to PAPB should be made. First, I should clarify the notion of behaving *differently*. PAPB requires behavior that is different in kind, rather than numerically different. So, PAPB is not necessarily satisfied by the possibility of a different token event. PAPB, like PAP, is motivated by the idea that a person cannot be considered blameworthy unless we could reasonably have expected better, nor praiseworthy unless we could have reasonably expected worse. Thus, PAPB requires not simply the possibility of a numerically different behavior event, an event that would be a wash, morally speaking, but behavior that is different in a morally significant way. If the agent is blameworthy for her behavior, PAPB requires the possibility of behavior for which she would not be blameworthy, and if the agent is praiseworthy, PAPB requires the possibility of behavior for which she would

not be praiseworthy.[107] So, the requirement of different possible behavior may be stronger than is immediately apparent from the statement of the principle above.

Second, I will defend a compatibilist version of PAPB only. Compatibilist versions of PAP typically make use of some sort of volitional event. They require that the person would have done otherwise, if he had so *chosen* or if he had so *decided* or if he had so *willed*. This is not the approach I favor for PAPB. The very point of PAPB is to take account of a third possibility, on which the agent does not act voluntarily or as the result of a volitional event. Instead, PAPB should tie the possibility of behaving differently to certain relevant differences in motivation or moral concern. The conception of blameworthiness developed in Chapter Two shows that blameworthiness for behavior requires that the behavior reveals that the agent's motivation to act in accordance with the relevant normative

[107] It might be suggested that PAPB should only require the possibility of behavior for which a person is *less* blameworthy (or for praiseworthiness, the possibility of less praiseworthy behavior). But consider the proposed counterfactual scenario, in which the agent behaves in a way for which she is less blameworthy. Applying PAPB to this counterfactual blameworthy person, there would have to be yet another alternative scenario in which she would be still less blameworthy, and so on *ad infinitum* or at least until we find an alternative in which she is not blameworthy at all. So, we should take PAPB to require, for blameworthiness, an alternative in which the person is not blameworthy for her behavior.

One might worry that this rules out the possibility of situations in which a person is blameworthy no matter what she tries to do. It seems as if there may be such situations, but we can make room for them if we suppose that what the agent is blameworthy for is getting into a situation in which she has only bad options. Her blameworthiness for creating or permitting that situation may transfer and make her blameworthy regardless of the fact that she chooses the best option once she is in it. If, on the other hand, she is not at fault for creating her dilemma, then I suspect that she will not be blameworthy for choosing what she reasonably concludes to be the least bad option.

considerations is weaker than we may reasonably expect of her.[108] So, we should expect that if the blameworthy agent were more motivated to act in accordance with normative considerations, if she were as concerned and motivated as we might reasonably expect, then her behavior would be relevantly different. With some slight changes, the same is true of praiseworthiness. If a person is praiseworthy for behavior, it is because that behavior shows that her motivation to act in accordance with normative considerations is stronger than we might reasonably expect. If her motivation were not this strong, then her behavior would be relevantly different. Therefore, the compatibilist version of PAPB should tie the possibility of behaving differently to relevant changes in the agent's motivation to act in accordance with reasons.

The compatibilist version of PAPB can be expressed most precisely as separate principles for blameworthiness and praiseworthiness:

[108] The idea of an agent's motivation to act in accordance with relevant normative considerations is somewhat awkward and arguably over-intellectualized. I have been tempted to rephrase the idea in simpler and more familiar terms, such as an agent's concern to do the right thing or the best thing. But I believe the phrase I have chosen allows some needed flexibility. I suspect that there are many circumstances in which there are relevant normative considerations that we should be concerned about and motivated by, but there is no right thing to do. I also suspect that there are situations in which we might praise, or withhold blame, based on a person's concern to do *some* good, even if the person lacks a desire to do *the* right thing or *the* best thing. Perhaps I should give more to charity and perhaps if I were less selfish I would, but I might still deserve some credit for being motivated to give as much as I do. I do not know how much is the right amount to give, and in the end, that question does not drive my decisions about how much to give. I give the amount that I do because I want to do some good, but I am not so strongly motivated to do the best thing that I possibly can. I recognize that there are reasons to give, and I act on them, but I choose not to be concerned about doing the best thing I could do. I do not know whether I should be considered praiseworthy or blameworthy for my giving, but in evaluating my blameworthiness or praiseworthiness, it should be relevant that I am motivated by normative considerations, even if I am not always motivated by a concern to do the right thing or the best thing.

> If a person is blameworthy for behavior, then if she were more motivated to act in accordance with normative considerations, she would have behaved in some other way for which she would not be blameworthy.
>
> If a person is praiseworthy for behavior, then if she had been less motivated to act in accordance with normative considerations, she would have behaved in some other way for which she would not be praiseworthy.

We should now be in a position to see what can be said for this proposal.

B. The Argument for PAPB:

Now let us consider the basis for believing PAPB states a necessary condition for responsibility. An argument for PAPB, as applied to blameworthiness, can be made out if we accept the following premise:

> If A is blameworthy for his behavior, then, during the relevant period prior to his behavior, A's motivation to act in accordance with normative considerations was weaker than we may reasonably expect of him.

I will have more to say about this premise below, but let me first say how it can be used to support PAPB. If we accept it, then a second premise naturally follows through contraposition:

> If, during the relevant period prior to his behavior, A's motivation to act in accordance with normative considerations had been as strong as we could reasonably expect, then A would not be blameworthy for his behavior.

From these premises, I infer the following:

> If A is blameworthy for his behavior, then if, during the relevant period prior to his behavior, A's motivation to act in accordance with normative considerations had been as strong as we could

> reasonably expect, then A would not be blameworthy for his behavior.

This is enough to establish PAPB as applied to blameworthiness, because, as I shall argue below, if in one scenario A would be blameworthy and in the second scenario he would not be, then his behavior in the two scenarios is different in kind. In the two scenarios, I shall argue, A behaves in a different way. With the necessary changes, the same premises could be used to establish PAPB as applied to praiseworthiness.

The argument for PAPB only gets going if we accept my first premise. I take that premise to follow from the conception of blameworthiness and praiseworthiness defended in Chapter Two. There, I argued that blameworthiness for behavior requires that the behavior in question reveals an agent's weak concern for normative considerations, and that praiseworthiness for behavior requires that the behavior in question reveals an agent's strong concern for normative considerations. The strength of an agent's motivation to act in accordance with normative considerations is revealed by her behavior only if it figures among the causes or explanations of that behavior. So, if a person is blameworthy for her behavior occurring at a particular time, then over some relevant period prior to that behavior, her concern or motivation to act on moral reasons was not as strong as we might reasonably expect.

The argument follows naturally from this idea. If we take an actually blameworthy agent, and then imagine her to be strongly and consistently motivated by normative reasons, then her behavior could no longer be of a kind for which she would be blameworthy. In short, altering the features of her

will that are required to make her behavior blameworthy is sufficient to ensure that her behavior would be relevantly different.[109] Similarly, praiseworthiness for behavior depends on behavior that reveals the agent's strong concern for reasons. If we take a praiseworthy agent and imagine her to be less concerned for reasons, her behavior could no longer be caused or explained by the strength of her concern for reasons. Thus her behavior would be different. It would no longer be the sort for which she could be praiseworthy. In either case, PAPB follows from the fact that moral responsibility for behavior depends upon the strength of the agent's motivation to act in accordance with normative considerations.

Contrast the foregoing rationale for PAPB from one that might be offered in support of PAP. It is sometimes supposed that alternate possibilities requirements get their support from an idea that blame must be avoidable (*i.e.*, that it would be unfair to blame a person unless she could have done something to avoid blame). That is not the basis of my argument for PAPB. I am not certain what to make of the idea that blame must be avoidable as a rationale for PAPB. For one, the avoidability intuition does not explain why any requirement like PAP or PAPB should apply to praiseworthiness. One could certainly believe that a person only deserves praise for an action that he might have avoided. But to praise where an action could not have been

[109] Alternatively, if we imagine this agent to be as strongly and consistently concerned about normative considerations as we could possibly demand and yet this makes no difference to her behavior, then this would show that, in the actual world, the weakness of her concern for reasons was not among the causes for her behavior and that she was not in fact blameworthy.

avoided is not obviously a wrong to the recipient of the praise (or anyone else) in the same way that blaming can be a wrong to the recipient of blame. Perhaps, in some indirect way, undeserved praise could cause harm to the recipient, as spoiling or coddling can cause harm, but it would be difficult to make a compelling case that undeserved praise is always wrongful to the recipient. As I understand it, the force of the avoidability intuition comes from this additional element: that it is wrong or unfair to the recipient unless she had some opportunity to avoid it. Without this additional element, the idea that praise is deserved only when it could be avoided seems to be just another way of stating that responsibility requires alternate possibilities.

I also do not rely on the avoidability intuition because I am not convinced that the compatibilist can adequately satisfy this intuition. Every compatibilist about determinism and moral responsibility must concede that blame is not avoidable in a fundamental and important sense. He must concede that blame is "avoidable" only in a sophisticated, philosophical sense, typically expressed through counterfactuals.[110] Some compatibilists might have a good story to tell about why this counterfactual sort of avoidability should do enough to satisfy our concerns about the fairness of blaming, but it becomes more difficult for the compatibilist to claim avoidability as a first principle that can support PAP or PAPB.

The intuitions stated at the outset of this Chapter provide another

[110] For example, PAPB supports the following position: blame can always be avoided simply by being better motivated. If one deserves blame, then if one had only been more motivated to do the right thing, or more concerned to act in accordance with normative reasons, one could have thereby avoided being blameworthy.

source of support for PAPB. I have claimed that in order for a person to be blameworthy for her behavior, it must be reasonable to expect that she behave better. This is not the same as saying that she could have done otherwise or that blame must be avoidable, only that we could have reasonably expected something better. The sense of "expect" here is normative. We might also say that it must be reasonable to have *demanded* or *asked* better of her.

What could we have reasonably expected of a person who acts badly? In some cases, it is simple: we could have expected or demanded that she do the opposite of whatever it is she is blameworthy for doing. For example, we may reasonably expect that a blameworthy liar should have told the truth. But it will not always be the case that a person could have done the opposite of what she did. In Chapter One, we saw that we may have the ability and opportunity to do something even though luck or other factors may prevent us from succeeding in some cases in which we exercise the ability. For example, a lifeguard may have the ability and opportunity to rescue even if there is a chance that he might get a cramp while trying or even if there is a chance that the rescuee will choke and drown shortly before the lifeguard can get there. A lifeguard who does not even attempt to rescue can be blameworthy even if, had he tried, one of these risks would have prevented him from succeeding. We can still demand better of this lifeguard because we can demand that he should have tried to rescue.

Of course, the alternative of trying is not always available. Frankfurt

cases appear to show that, in some cases at least, we cannot even expect the blameworthy person to get as far as trying. If we concede that a Frankfurtian intervener could step in before the agent even manages any effort to do otherwise, then we must concede that we cannot even demand that the blameworthy person try. What we can demand in every case is that she had been more concerned or motivated to act in accordance with the relevant normative considerations. If she were, then her behavior would have been different, regardless of whether or not bad luck or a Frankfurtian intervener prevents her from successfully doing what she ought to do. Her behavior would have been different because this increased motivation or concern would figure in the explanation of that behavior.

Could we also expect or demand that a person have better abilities or opportunities? In some normal cases, we will feel that a person is blameworthy because she failed to cultivate better abilities, abilities that would have allowed her to overcome the obstacles to doing what he ought. If John the lifeguard is out of shape, he may be blamed for failing to rescue even if he tries. If an alcoholic puts herself in circumstances where she will be tempted to drink, she may be blameworthy for drinking. But this does not show that blame can be deserved simply because we may demand better abilities or opportunities. We may demand these things only to the extent that a person would be able to acquire them, *if* she had been more motivated to do so. In other words, if we blame in these circumstances, we assume that if the person had been more motivated by the relevant normative considerations, she would

have cultivated stronger abilities or put herself in better circumstances. Thus, even in these cases, the assessment of blameworthiness is ultimately grounded on a demand that the person should have been better motivated, that she should have been more motivated by reasons. The same is true where a person's behavior would have likely been better if she had cultivated better specific habits or dispositions. Ultimately, it is her failure to be motivated by the reasons for cultivating these more specific habits that permits the assessment of blameworthiness. At bottom then, all that we can reasonably demand is that the person be motivated to act in accordance with relevant normative considerations. If she is so motivated, then she will do what we can reasonably demand to improve herself and her situation. Thus, if she is blameworthy for her behavior, then it must be that her behavior reflects, in one way or another, the weakness of her motivation to act in accordance with normative considerations.

Now consider praiseworthiness. For a person to be morally praiseworthy for her behavior, that behavior must reflect a concern for reasons that meets or exceeds our reasonable expectations. In such a case, it is typically quite reasonable to expect—now in a more predictive than normative sense—less of her. We understand that human beings are imperfect moral agents. They are subject to temptations of all sorts. A person will often show concern that meets or exceeds our expectations where she does not give in to a normal and understandable temptation, either because she has a stronger desire to do what she ought or because, over time, she has cared enough

about doing the right thing to develop good habits and suppress bad desires. But we will not consider her to be morally praiseworthy if, instead, her behavior merely reflects good luck, including the strength of her innate abilities or the good circumstances in which she finds herself.

Whether she will deserve praise when her behavior reflects a natural tendency to be kind, for example, is a more subtle question. I believe that it may, but that the praise should be qualified if that tendency is not curbed by a general concern to do the right thing. We will want to praise and encourage a tendency to be kind (or toward some other virtuous disposition) when it is exhibited in appropriate situations, but if a person is naturally disposed to be kind no matter what the situation, we may see that as a fault rather than an occasion for praise. It is hard to doubt that there are situations in which simple kindness is inappropriate, and in which it is better to be firm, critical, or even harsh. In extreme cases, where one is dealing with an abusive person, uncurbed kindness may result in serious harm to oneself and others as well. Thus, in general, I believe that a person deserves praise only to the extent her behavior reveals the strength of her concern to do the right thing.

There is only one other point in this argument that may deserve further comment. The premises I have defended so far support the conclusion that if a person is blameworthy for her behavior, she would not have been blameworthy for her behavior if she had been more strongly motivated by normative considerations. In contrast, PAPB provides that if a person is blameworthy for her behavior, then she would have behaved in a different way

if she had been more strongly motivated. Thus, PAPB follows from my claims only if we accept that the actual behavior resulting from one's actual motives and the hypothetical behavior resulting from the hypothetical motives are different ways of behaving, *i.e.*, different in kind. This additional premise reflects my view that it makes sense to sort behavior into kinds based on whether or not a person is blameworthy, praiseworthy or neither for that behavior. My general view is that kinds, or at least behavioral kinds, are as varied as our interests, and that we may meaningfully sort behavior—just as we may sort other things—based on any properties that we care about. For our purposes, behavior for which a person is blameworthy is very different than similar behavior for which a person could not possibly be responsible. For example, I believe that whipping a donkey because one wants to see it in pain is different behavior than whipping a donkey because one reasonably believes that it is the only way to get it to move to a place of safety. I also believe that each of these behaviors is different than whipping a donkey because one is an automaton under the thoroughgoing control of an evil genius.

If one does not share my permissive view of behavioral kinds, there are still adequate grounds for insisting that it is meaningful and appropriate to distinguish kinds of behavior in terms of the attitudes or mental states that cause the behavior. We very often sort behavior according to the motives and mental causes. The criminal law is instructive here. Systems of criminal law with origins in the British common law, including our own, rely extensively on

the mental states that explain and motivate that behavior in order to define categories of behavior. First, to count as any sort of crime at all, the behavior at issue must be voluntary, in the weak sense identified above. In addition, few crimes are defined simply in terms of overt acts, but in terms of the acts together with the mental state with which these acts are carried out. For example, in the traditional common law, homicide is murder and houseburning is arson only if the requisite conduct is performed with "malice." If homicide is committed by the very same means and physical movements but in the heat of passion induced by one of a few recognized forms of provocation, then it is manslaughter instead of murder. If committed by the same means, but as part of a reasonable response to a life-threatening attack, then it is not a crime at all. In jurisdictions that have modified the traditional common law, the precise lines between the categories have shifted a bit, but the general point remains true. Criminal acts and omissions are, like blameworthy and praiseworthy behavior, defined and categorized at least partly in terms of the mental states that motivate them. It would defy common sense to insist that criminal behavior is not different in kind than similar non-criminal behavior or that murder is no different than manslaughter. I find it similarly incredible to claim that blameworthy behavior is not different in kind from morally neutral or praiseworthy, but otherwise similar, behavior.

Based on the foregoing considerations, there is a clear argument for PAPB that depends on the fact that desert of praise and blame requires behavior that reveals the quality of a person's motives. If a person is

praiseworthy for her behavior then that behavior is caused by the strength of her basic moral desires. It follows then that if her desires were sufficiently different in this respect, then even if her behavior would be otherwise the same, it would not reveal the strength of her moral motivation. Thus, it would not be behavior for which she could be praiseworthy. Similarly, if a person is blameworthy for her behavior, then her behavior was caused by the weakness of her moral desires. So, if her desires had been sufficiently different, then her behavior would have been different. It would not be behavior for which she could be blameworthy.

C. PAPB and Frankfurt Cases:

PAPB is designed with Frankfurt-type cases in mind, and it concedes that a person could be morally responsible without any relevant possibility of *doing* otherwise. So, it may be no great surprise that Frankfurt cases cannot be used to refute it. Nonetheless, this is a fact that should be shown.

Frankfurt's main case against PAP is given in the following counterexample:

> Suppose someone—Black, let us say—wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something *other* than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. Whatever Jones's initial preferences and

inclinations, then, Black will have his way.[111]

It might be more accurate to call this a schema for a counterexample. As Frankfurt says, the details of just how Black can predict what Jones will do and, if necessary, successfully intervene can be filled in various ways. And other philosophers have crafted different versions of this general scenario in an effort to meet every possible concern and objection. For the purposes of testing PAPB as a part of a compatibilist theory of responsibility, we need only focus on two broad kinds of Frankfurt cases. The first is the general family of cases that get the most attention, those in which Black uses a device to manipulate Jones's brain or directly controls his mind. The second is rarely discussed but still faithful to Frankfurt's description here. These are cases in which Black does not interfere with Jones's power to decide what to do when he intervenes. Black simply ups the ante sufficiently to, as Frankfurt puts it, "ensure that Jones decides to do, and that he does do, what Black wants him to do."

Let us begin with a direct manipulation case, in which Black has some sort of brain or mind control over Jones. Suppose that Jones and Black are jurors in Red's criminal trial. The case against Red is very weak, but most of the jurors have an unreasonable prejudice against Red. Nearly all of the jurors are overlooking the glaring holes in the case and will vote to convict on the basis of prejudice alone. Black in particular is so determined to see Red convicted that he will take extraordinary measures to ensure a unanimous

[111] Frankfurt (1969) p.835

guilty verdict. Jones is alone among the jurors in feeling conflicted. He does not want to disagree with his fellow jurors, but he has much more than a reasonable doubt about Red's guilt. Black hopes that a little cajoling and peer pressure during deliberations are all that will be needed to prompt Jones to vote to convict, but if not, he is prepared to use a device to manipulate Jones's brain long enough to have Jones's vote cast for Red's conviction. If we suppose that Black can do this, then we might as well imagine that Black can also monitor Jones's brain processes and, from them, accurately predict what Jones will do. We may therefore assume that Black will intervene effectively in any situation in which Jones would vote to acquit and, thus, that Jones's vote will be cast to convict no matter what. But in fact Black never needs to intervene. During deliberations in the jury room Jones expresses some misgivings, but when he looks around the table and sees everyone glaring at him, he caves. Unwilling to stand up to them, he joins them in voting for a conviction regardless of his belief that Red has not been shown to be guilty.

It should be quite clear that Jones is blameworthy for voting to convict. He has more than a reasonable doubt. So, Jones has both a legal and a moral duty to vote to acquit. Red's liberty is at stake, after all. Let us also grant that Jones could not *do* otherwise.[112] If he had been inclined to hold out against conviction, Black would have intervened and taken control of Jones's

[112] We could quibble a bit about this. Perhaps he could have argued a bit more. Perhaps he could have begun to form an intention to vote to acquit. If so, perhaps the example could be reformed to rule as much as possible of this out by having Black intervene earlier. Then the search would be on for something else, however slight, that Jones might have done before Black intervened. But I am not defending PAP here, and I need not go down this road.

brain and, thus, his body. This does not contradict PAPB. Though Jones could not have done otherwise, his behavior could have been significantly different. If Jones had simply been a better person, if he simply had been more strongly motivated to do his duty, then he would have had adequate motivation to hold out. If he had been so motivated, it is true that Black would have intervened and ensured that Jones's vote is cast to convict, but Jones's behavior still would have been different. If Black intervenes in this sort of way, then *Jones* does not vote to convict. He does not *do* anything at all.[113] Black's intervention breaks the connection between Jones's beliefs and desires—his will—and his behavior. All that happens in the counterfactual scenario is that Jones's body is made to move in ways that appear as though he votes to convict. This is quite different than what Jones actually does. It is certainly not behavior for which Jones is blameworthy. So, this case poses no threat to PAPB.

Next suppose that the manipulation of Jones's brain or his mind is not so direct. Suppose that Black would intervene not by directly causing the behavior he wants, but by giving Jones new desires or by increasing the

[113] As noted above, there might be a weak sense in which Jones does vote. If Black's intervention interferes with Jones's awareness of his own behavior, he may well want to know "what he did" while he was out of control. Or if he remains aware but unable to control his behavior, then he may watch in horror at what he finds himself "doing." If one prefers to speak this way, then we need only note that though there is still something Jones does in the counterfactual scenario, it is dramatically different from what he does in the actual scenario. In the actual scenario he blamefully and voluntarily votes for conviction. In the counterfactual scenario, he blamelessly and non-voluntarily is manipulated into voting for conviction. But I have my doubts about saying even this. I am inclined to say that voting is essentially voluntary, and so that Jones only appears to vote in the counterfactual scenario. So, if one insists that Jones does anything in the counterfactual scenario, I think one can only say that he non-voluntarily makes other motions or words that appear to be voting for conviction.

strength of certain existing desires. Perhaps we can imagine such a case. In the actual case, Jones is actually motivated to vote for conviction because his fear and anxiety over confronting his fellow jurors is stronger than his concern to do his duty. Suppose that it became clear that Jones would not be sufficiently moved by this anxiety, Black might intervene by manipulating Jones's brain chemistry in such a way that his feelings of anxiety become overwhelming. Perhaps one will claim that if this works to motivate Jones to vote to convict, then voting to convict is still something that Jones *does*. I disagree. The idea that this new, heightened anxiety and the desires that come with it are *Jones's* is, to my mind, perverse. Even if we should say that Black's intervention gives rise to new desires in Jones, we should not say that these are Jones's desires.[114] So, the intervention still breaks the connection between Jones's desires and his behavior and, as a result, prevents him from doing anything voluntarily. Therefore, I am not inclined to suppose that this move permits us to say that *Jones* votes to convict in the scenario in which Black intervenes.

[114] As this makes clear, I am inclined to suppose that some sort of ownership requirement applies to one's motives and desires. I am not prepared to make this idea precise here, but I do not think I need to. Whatever concept is defended would have to be sophisticated. People can intentionally cause each other to have desires, and the created desires very often become a part of the agent's desires. Education and advertising make this clear. In many cases, the bearer of the created desire need not knowingly and affirmatively consent before the desire becomes a part of his will. However, there is likely a role for something like consent as well. Medication may change our desires. If the medication is secretly given to us without our consent, there is some reason to suppose that the new desires interfere with rather than change our wills, but if it is taken with informed consent such an argument becomes more tenuous. In any case, without being able to give a full set of criteria, I think it is safe to say that if, without consent, one person interferes with desires of another as directly and as effectively as we must imagine that Black does, the created desires are not the second person's desires, at least not immediately.

Even if this were not so, Jones would not be blameworthy in the counterfactual scenario, and thus, his behavior would have been different. In the counterfactual scenario, Jones's behavior is caused by an artificially stimulated and overwhelming desire, not by the weakness of his moral motivation. By hypothesis, in the counterfactual scenario, Jones is as concerned and motivated to do the right thing as we could reasonably expect. This strong moral concern is simply overwhelmed by an artificially induced pathological desire to the contrary. Thus, however we should characterize his behavior in the counterfactual scenario, it cannot support an attribution of blameworthiness. Jones either does nothing at all because his behavior is not the product of his will, or he does something quite different than his actual behavior.

While the literature is dominated by cases in which Black would intervene by using some extraordinary technology or magic to meddle with Jones's brain, Frankfurt suggests that there are other ways of getting his schema to work. Frankfurt himself proposes that Black might intervene by "pronounc[ing] a terrible threat, and in this way both force Jones[] to perform the desired action and prevent him from performing a forbidden one."[115] This method of intervening has the advantage of supporting some of the other things Frankfurt says about Black and Jones. In particular, Frankfurt says that if Black must intervene, he may take steps to "ensure that Jones *decides* to

[115] Frankfurt (1969) p.835.

do, and that he *does do*, what [Black] wants him to do."[116] As I have just argued, in the typical brain manipulation cases, this description is not honored. In those versions, if Black intervenes, then Jones makes no decisions, and strictly speaking, he does nothing. So, we should consider a case in which Black intervenes with threats and other coercive techniques. If nothing else, consideration of such a case will provide another illustration of how PAPB works.

Suppose that Jones is the single breadwinner in a poor family. His frail son, Tiny Tom, is suffering from a persistent and worrying cough, and the gas company is threatening to turn off the heat. Jones has just cashed his paycheck, and he is headed home, hoping he has enough money to keep the heat on and to buy a modest but hearty Christmas meal for the family. Black, however, wants to get drunk, and knowing that Jones enjoys a drink when he can, Black thinks he might just be able to get Jones to buy a few rounds. So, when Black sees Jones coming down the street, he proposes they stop in at the pub and spend just a little bit of his pay to celebrate. The thought of putting his troubles out of mind for a moment appeals to Jones, and soon enough, amidst the fog of ale, he has spent nearly his entire pay on drinks for the two of them. Jones never realizes just how desperate Black was. Had Jones turned him down, Black was prepared to threaten Jones at gunpoint. He would have marched Jones right into the bar and ordered round after round, forcing Jones to pay. So, either way Jones would have spent his check

[116] Frankfurt (1969) p.835 (emphasis added).

on drinks rather than on basic necessities for his family. But in the actual case, Jones is blameworthy because he decided to do it for his own reasons, without being threatened.

It may be clear why this sort of Frankfurt case has been of less interest in the debate over the principle of alternate possibilities. Few defenders of PAP will concede that Black could really rule out the possibility of doing otherwise by intervening in this way. After all, even if Black raises the stakes to the point where no reasonable man would resist, Jones still has a choice. He just has a bad choice. Though it might be unreasonable to do so, Jones could always force Black's hand and risk being shot. So, in a case like this, a defender of PAP might concede that Jones probably would do what Black wants, but not that Jones could not, strictly speaking, do otherwise.

But PAPB works somewhat differently PAP. In the counterfactual scenarios relevant to PAPB, Black can virtually ensure that Jones goes to the bar. PAPB asks us to consider a counterfactual scenario in which Jones is strongly motivated to do the right thing. If we assume that this means he is strongly motivated to provide for his family, then it may be appropriate to assume that he also would not risk being shot when they are depending on him not just for this week's pay but for years to come. So, perhaps we should concede that Black can raise the stakes high enough to ensure that Jones will spend his check at the bar. In this sort of case, Jones still *does* something in the counterfactual scenario. He still goes to the bar and spends his pay. These acts are voluntary in the weak sense identified above. But that is not to

say that he does the same thing that he does in the actual scenario. There is a great difference between irresponsibly frittering away the family's money in order to forget about reality and spending that money solely because a person has a gun to his back. To appreciate the difference, one need only imagine how the family should feel in each of these scenarios. In the actual scenario, Jones is blameworthy. In the counterfactual scenario, he is not. He is not responsible at all, or perhaps he might even be praiseworthy for his calm and reasonable response to the threat. Thus, this case again illustrates the central point of PAPB. If Jones's moral motivation was stronger, then his behavior would have been different.

In sum, Frankfurt cases do not present a counterexample to PAPB. Black's counterfactual intervention can ensure that Jones's counterfactual behavior is *similar* in some ways to his actual behavior, but that intervention also ensures that his behavior is different in the relevant respects. By hypothesis, in the counterfactual scenario Jones is strongly motivated by the relevant reasons, and that entails that he is not blameworthy for whatever behavior Black manages to bring about. Whatever similarities there are between his actual and counterfactual behavior, they are quite different in this respect.

D. The Relevance of Behaving Differently

One might suppose that PAPB is true but uninteresting. Some of Frankfurt's supporters have been willing to concede that Frankfurt scenarios cannot rule out every sort of alternate possibility, but they have suggested that

these ineliminable alternate possibilities are insignificant and irrelevant to moral responsibility. John Martin Fischer acknowledges that Frankfurt examples leave open alternatives. Yet he claims to have arguments that "come extremely close to establishing that alternative possibilities are not required for moral responsibility."[117] Similarly Michael McKenna and Derk Pereboom have attempted to buttress Fischer's claims by identifying specific criteria that distinguish relevant and irrelevant alternate possibilities. Generally, the targets of their arguments are attempts to save PAP, not attempts to defend PAPB. Nonetheless it is worth considering what these arguments might show about PAPB. I shall argue that they do not show that PAPB is false.

However, their arguments may also evoke a deeper concern. These arguments may be seen as giving voice to doubts about the significance of any sorts of alternatives left open in Frankfurt scenarios. After all, there is no shortage of necessary conditions for responsibility (*e.g.*, In order for A to be blameworthy for X, there must be something rather than nothing at all; in order for A to be morally blameworthy for X, there must be some moral truths). Very few of these necessary conditions tell us anything interesting about responsibility in particular. So, one may wonder whether PAPB states a condition that really shows us much about moral responsibility. I will address this concern with a brief discussion of the significance of the principle and the alternatives it requires.

[117] Fischer (1994) p.147.

Fischer characterizes those who emphasize the importance of the alternatives left open in Frankfurt scenarios as "flicker of freedom" theorists.[118] He supposes that those who insist on the importance of these alternatives must be emphasizing the existence of a small flicker of freedom—such as the freedom to φ willingly or to φ unwillingly—and arguing that such freedom provides an adequate basis for moral responsibility. Fischer's use of "freedom" shows that his point is not aimed squarely at my position. Nonetheless, Fischer's broad description of one version of the "flicker of freedom strategy" sounds a good deal like the position I have developed here. This version, Fischer says, emphasizes the difference between Jones acting on his own and Jones behaving as the result of Black's intervention, and it insists that the existence of these alternatives are crucial to moral responsibility.[119] Fischer does little more to develop this precise line of reasoning. He does not directly say, for example, why, on the line of reasoning he is imagining, these sorts of alternatives are crucial or whether they constitute a sufficient or merely a necessary condition for responsibility. Nonetheless, Fischer's rough sketch is reminiscent enough of my position that we ought to consider why he rejects it.

Regarding such alternatives, and indeed regarding any variant of the "flicker of freedom" strategy, Fischer says that his "basic worry" is that Jones's alternative to acting on his own "is not sufficiently *robust* to ground the relevant

[118] *Id.* p.134.

[119] *Id.* p.139.

attributions of moral responsibility."[120] Fischer restates this worry in a number of ways. In places, he describes Jones's alternative to acting on his own desires as "exiguous," and in others, he describes it as "etiolated." A quick look in the dictionary confirms that these words mean "thin," "weak," and, one supposes, "not robust." This is impressive vocabulary, but is it an argument? The use of these terms is plainly metaphorical. So, we should expect some explanation of the kinds of features that make alternatives either "robust" or "exiguous." Even if we were to concede that the alternative required by PAPB is "etiolated," we need some explanation of why this matters. Unfortunately, Fischer does not directly explain what makes alternatives robust or exiguous.

We can glean some sense of what Fischer's concern is from certain rhetorical questions he asks. The "too thin" intuition is developed by asking, in various ways, "How could adding a set of alternatives in which Jones does *not* act freely make it the case that he *actually* acts freely?"[121] As he restates and explains this question, Fischer repeatedly suggests that alternate possibilities, if relevant, must "ground" attributions of moral responsibility, and he complains that it could not be "in virtue of" these sorts of alternate possibilities that a person is morally responsible.[122] None of these phrases ("make it the case," "ground," and "in virtue of") is precise, but they all suggest that Fischer is working with an expectation that alternatives, if relevant, must constitute a *sufficient* condition for moral responsibility. We would not normally say that A

[120] *Id.* p.140.

[121] *Id.* p.141-2.

[122] *Id.*

makes B the case unless A is sufficient to bring about B or perhaps, if not sufficient on its own, an especially salient part of a sufficient condition that is otherwise satisfied. We would not normally use any of these phrases unless we expected the existence of alternatives to be the key part of the explanation for moral responsibility.

PAPB is stated as a necessary condition. Thus, Fischer needs to show that the alternate possibilities left in Frankfurt cases, which PAPB exploits, are not *required* for moral responsibility, but he seems to be more focused on whether or not they are *enough* for moral responsibility. For our purposes then, Fischer is asking the wrong question. The question is not, "How does adding alternatives make it the case that a person is responsible?" It should instead be, "How would the absence of alternatives make it the case that a person is not responsible?" The argument for PAPB makes the answer to this question clear. If a person's behavior would remain the same no matter how much the strength of his moral motivation changed, then his behavior does not reveal anything relevant about his moral motivation and thus he could not be morally responsible for it. In any case, even if PAPB did state a sufficient condition, it would be wrong to suppose that it is the alternatives themselves that somehow "ground" moral responsibility. The alternatives might be said to be like symptoms of the relevance of moral motivation. Their existence follows from the fact that behavior for which a person is responsible is behavior that reveals the strength of her moral motivation, not vice versa. And it is this latter fact that "grounds" moral responsibility.

Perhaps McKenna's or Pereboom's elaborations of Fischer's point will show us why robustness should concern us. McKenna identifies two specific criteria for a robust, or relevant, alternative. He writes that "a robust alternative requires two conditions":

> (1) The alternative is morally significant. It would tell us something (different than what we are told in the actual world) about the moral quality of the agent's conduct were she to have so acted in this alternative scenario.
>
> (2) The alternative has to be within the control of the agent.[123]

I have already argued that the alternatives required by PAPB are morally significant, and McKenna is prepared to concede this point. He acknowledges that acting as the result of Frankfurt-type intervention does not reveal the same thing about one's will as acting for one's own reasons. "Hence," he writes, "the alternatives [left open in Frankfurt cases] are morally significant; they bear importantly different moral properties with regard to an evaluation of [the agent's] conduct."[124] So, the interesting issue here is the second criterion, control.

Control is not an easy concept to define. To understand what McKenna means by control, we need to look to his argument that "prior sign" Frankfurt cases can be used to rule out all alternatives "within the control of the agent." Prior sign cases are a class of Frankfurt-style cases in which Jones (or his equivalent) reliably and involuntarily displays some sign indicating what he will do before even he knows what he will do. Sometimes these prior signs are

---

[123] McKenna (2003) p. 204

[124] *Id.*

imagined to be simply extremely reliable poker "tells"—little bits of involuntary behavior that give Black advanced information about what Jones will decide. But it is also sometimes supposed that Black, a super-neuroscientist, has identified a brain pattern or signal that occurs just before Jones makes a decision and indicates what he will decide. McKenna describes such a case.

In McKenna's case, Leslie (taking the place of Jones) has a plan to shoot Pat with a squirt gun, and Daphne (taking the place of Black) wants to make sure that she does it. Daphne has been monitoring Leslie's brain and has discovered that if Leslie is going to decide to shoot Pat, she will inevitably display a particular brain signal shortly beforehand. So, Daphne knows that if Leslie does not display the brain signal by a certain point she is not going to decide to shoot Pat, in which case Leslie will intervene by making use of a "gizmo" that will cause Leslie to shoot Pat. In the actual case, Leslie displays the brain signal and shoots Pat on her own. (To ensure the case has some moral significance, we might imagine that Leslie knows the water in the squirt gun will ruin Pat's treasured, one-of-a-kind outfit and make Pat very upset.)

McKenna makes the following claim about the relevance of the counterfactual alternative in which Daphne intervenes:

> This alternative cannot be relevant to evaluating Leslie's free will and moral responsibility with regard to her conduct in the *actual* world, the world in which she does shoot Pat on her own. Why? Because such an alternative, arising as it does prior to the locus of Leslie's free will, was not within the scope of *Leslie's* control.[125]

[125] *Id.*

This description gives us our best hint as to what McKenna means by "within the agent's control," but we need to unpack this terminology to understand it. First, we have the suggestion that the alternative should be said to arise at a particular time. This is perhaps natural enough, but we should make it a bit more explicit. Presumably the alternative "arises" at the point where it diverges from the actual course of events. In this case, this might occur at the time when Leslie displays the brain signal, for it is that point that we know the counterfactual scenario is supposed to diverge from the actual world.[126] (It might be a bit more accurate to say this is the point at which the alternative becomes closed off.) Second, and more boldly, we have not only the assumption that there is something meaningfully referred to as "Leslie's free will" but also the assumption that it has a "locus," apparently a temporal one. McKenna is clearly suggesting that this temporal locus is something that occurs after the time at which Leslie displays the relevant brain signal. Because we are to suppose that the brain signal is a reliable indicator that, moments later, Leslie will decide to shoot Pat, and because no other candidate events are mentioned, it is reasonable to infer that McKenna is supposing that the moment at which Leslie *decides* to shoot Pat is the "locus of her free will." So, I believe that the following captures McKenna's claim in

[126] Note that, if we assume determinism, the idea that an alternative arises at this point, or at any other nearby point, may be a bit more controversial than I have suggested. Unless we suppose that minor, localized miracles are the proper way to evaluate counterfactuals in a deterministic world, then there is reason to suppose that in a deterministic world every counterfactual course of events diverges at the starting point of the universe (if there is one) and at every point since. But this point raises issues far outside the scope of this dissertation.

plain terms: any alternative left in this sort of Frankfurt case is not within the agent's control because the matter of which alternative will occur is already settled before the agent decides what to do.

In my view, the conception of control at work here reflects the grip of a fetish for decisions. I doubt that decisions have anything like the significance that philosophers often attach to them. For one, though it is true to say that we "choose" or "decide" to do pretty much anything that we voluntarily do, our voluntary behavior is very rarely preceded by any noticeable event of choosing or deciding. When I am driving, every minor adjustment in steering, in braking and in accelerating that I make is something that I have decided or chosen to do, but I make these adjustments without any conscious thought and without any awareness of choosing or deciding. When I am playing sports, I choose whether to pass or take a shot, whether to sprint downfield or backpedal slowly, whether to dodge left or right. But I do these things spontaneously, without any conscious deliberation or decision. Indeed, if I am to have any chance of playing well, I must simply react to what I see on the field and not deliberate and make conscious decisions. As a result, although it is perfectly correct to say that I choose to do everything that I do in these contexts, I find it highly suspect to postulate countless, imperceptible volitional events that precede each bit of my voluntary behavior.[127] Instead, in these contexts at

[127] Mele (2006) similarly recognizes that many intentions (followed by intentional actions) are formed without a decision. Mele offers the example of intentionally unlocking his office door: "since I am in the habit of unlocking my door in the morning and conditions . . . were normal, nothing called for a *decision* to unlock it. . . . [G]iven the routine nature of my conduct, there is no need to posit an act of intention formation in this case. My intention to

least, it makes more sense to suppose that talk of one's "choosing to x" or of "making a decision to x" is simply a way of saying that one did x voluntarily or, perhaps, intentionally.

Of course, there are those times when we are conscious of deliberating and when we have the conscious experience of making a decision prior to our action, but even in these cases the significance of these moments may be more epistemic, than metaphysical. In these cases, it typically seems as though any of a range of alternatives is open to us, that nothing prevents us from choosing any one of them, and then we know what it is that we will do. It is fairly natural to suppose that something quite special has happened in that moment, that we have exercised some faculty of choice, that a switch has been flipped in the future's garden of forking paths—provisionally closing off paths that were entirely open to us a moment ago. It is a quaint and attractive idea, but not necessarily a plausible or coherent one. There is certainly something wrong with this idea if the world is deterministic. In a deterministic world, all but one of those paths are, in an important sense, already closed. We simply do not know which one remains open. Therefore, in such a world, it would be quite plausible that what is special about these moments of conscious decision is not that they are moments in which we pick a path, but

---

unlock the door may have been acquired without having been actively formed." Mele (2006) pp.14-15. Mele's example of routine, habitual action is a plausible one. My examples show that I would extend the point to a wide range of actions that are less routine and that might seem to involve more active decision making of a kind.

moments in which we come to learn which path we are on.[128]

In any case, we constantly rule out courses of action without making any decision to do so and prior to making any decision about what to do. That does not show that it was not "up to us" or within the scope of our control to take those courses of action. I have frequently had, and I will often have, the option of going outside and running through the streets flailing my arms up and down while shouting at the top of my lungs. I have not done so, and I am unlikely to do so, though I cannot think of any time that I have decided against doing so. So far, this option has been ruled out simply by the fact that I have not considered it, by my tendencies to be shy and reserved to a certain degree, and/or by my interests in doing other things. But it would be strange to say that the option of doing so has not been within the scope of my control at any past time that I was deciding what to do. Surely, if I had been sufficiently inclined to run through the streets flailing my arms wildly, I could have done so. So, it has always been up to me, even if the alternative has been ruled out prior to each of my decisions about what to do next.

Similarly, in McKenna's prior sign case, which alternative occurs depends on what Leslie's desires and beliefs are prior to her decision. If she is mischievous and inconsiderate, she will shoot the gun voluntarily. If she is not, she will not, and she will instead be made to shoot it involuntarily. If her shooting did not depend on her desires in this way, then she could not be

[128] Admittedly, this is not all that happens. We are not merely observers gaining knowledge. In addition to forming a belief about what we will do, we form an intention.

responsible in the actual scenario. The dependence of the outcome on her desires in this way is an excellent *prima facie* case for supposing that whichever of these two alternatives occurs is up to Leslie, or put another way, within the scope of her control. The mere fact that a brain pattern shows what she will do prior to anyone's awareness of her decision, allowing Daphne to start her intervention earlier, seems to me to cast no doubt on her control over her actions.[129]

So, McKenna may be right that some form of control is a reasonable requirement for relevance of alternatives. But the alternatives required by PAPB and left open by Frankfurt cases can meet such a requirement if we simply adopt a more plausible notion of control: alternatives are within her control if they depend on which desires (within a relevant set that we could reasonably expect an agent to have) actually move the agent. If a particular set of events will occur (or will not occur) no matter what an agent desires, then it is plausible to say that they are not within the scope of her control. If some relevant change in that agent's desires and preferences would be adequate to make X happen rather than Y, then the matter of whether X or Y occurs is within that agent's control. This is an idea that is as simple and as

[129] McKenna's standard of control strikes me as problematic for another reason. McKenna contends that any alternative that "arises" prior to the agent's decision is not within the scope of his control. In his prior sign case, all alternatives are closed off prior to the agent's decision, and so, what the agent will do is settled prior to her decision. If this means that the alternatives are outside the scope of the agent's control, one must ask why it does not also mean that her actual behavior is outside the scope of her control. After all, just like the alternatives, the question of whether or not the actual behavior would occur was already settled prior to the "locus of her free will." There appears to be a double standard for control here.

familiar as any. It should not be confused for an *ideal* of self-control, an ideal that is in no way required for responsibility. Instead, it is a natural, minimal form of control that simply requires that what a person does depends on what he wants.

Moreover, this form of control actually does what McKenna intends. It rules out alternative possibilities that are actually irrelevant to moral responsibility. There are always countless counterfactual scenarios in which Jones (or Leslie or whoever) avoids doing what the counterfactual intervener wants. Perhaps a random contra-causal event occurs spontaneously wiping out Jones, Black or the entire world. Perhaps Black's gizmo has a technical failure. Perhaps the rapture begins. Or to take an example from Derk Pereboom, Jones might not engage in his blameworthy behavior, if instead he took a sip of coffee which unbeknownst to him, contains poison. These counterfactual possibilities are clearly irrelevant because whether or not they occur does not depend on relevant changes to the agent's desires. Other types of alternatives, those that do depend on the strength of an agent's moral motivation, are relevant to moral responsibility.

Derk Pereboom also identifies a notion of "robustness":

> For an alternative to be relevant per se to explaining an agent's moral responsibility for an action it must satisfy the following characterization: she could have willed something other than what she actually willed such that she understood that by wiling it she would thereby have been precluded

> from the moral responsibility she actually has for the action.[130]

I am not sure what it means to "will" something, but I suspect that deciding to do something is closely related, if not necessary and sufficient, to willing it. Given what I have just said about decisions, and given that PAPB focuses on alternatives in which one need not actually will or decide to do anything, I see no argument and no reason to follow Pereboom in assuming that an alternative is relevant only if it involves an event of "willing." But what Pereboom's condition adds to McKenna's, and what does require further discussion, is an epistemic element. Alternatives are only relevant, according to Pereboom, if the agent is aware of them and understands that they are such that the agent would, by choosing them, avoid the sort of moral responsibility she actually has.

Pereboom motivates his requirement by relying on claims about the avoidability of blame: "The main intuition underlying alternative possibility conditions is that if, for example, an agent is to be blameworthy for an action, it is crucial that she could have done something to avoid this blameworthiness."[131] Whether or not this intuition is needed to support PAP, it is not the intuition that I have invoked to support PAPB. As noted above, any compatibilist about determinism and responsibility must hold that, at a fundamental level, responsibility may not be avoidable. PAPB has two related bases. First, desert of praise and blame depends on the strength of a

[130] Pereboom (2001) p.26

[131] Pereboom (2005), Pereboom (2001) p.1.

person's moral motivation relative to our reasonable expectations. Second, a person cannot deserve praise or blame unless it is reasonable to have expected something different (worse or better, respectively) from them. It is not certain that either of these ideas requires or implies that the person could have avoided responsibility.

Suppose we grant Pereboom's point for the sake of argument. If one believes that blame must be avoidable, and if that is one's reason for requiring alternative possibilities, then an epistemic condition might seem to follow. Pereboom supposes that an alternative would not ensure or support the avoidability of blame unless the agent is aware that it would allow him to avoid responsibility.

This is a bit too quick. It may be true that the agent must have some belief about an alternative, but we should be careful not to impose a requirement of accurate knowledge about his alternatives. The avoidability intuition should be satisfied in at least some, if not all, cases where an alternative exists in which the agent would avoid blame (the "real alternative"), the agent has a mistaken belief about the existence of an alternative in which she would avoid blame (the "imagined alternative"), and if she were adequately motivated to bring about the imagined alternative, this would in fact bring about the real alternative. To illustrate, suppose that Fred's house is on fire and he believes that his wife is unconscious inside and that he could with little risk to himself get her out. Unmoved by the idea of his wife's peril, Fred stands by and watches the house burn. In fact, Fred's wife is not inside, but

unbeknownst to him, his daughter is. If Fred had tried to save his wife, he would have found his daughter and been able to rescue her instead. Is Fred blameworthy for failing to rescue his daughter? You'd better believe it.[132] Is the fact that he was not aware that there was an alternative in which he saves his daughter relevant? Ask his (soon to be ex-) wife, if you must. Mistaken though he was, Fred knew everything he needed to know about his alternatives in order to avoid blame.

So, even if we accept intuitions requiring the avoidability of blame, we should not accept any requirement of accurate knowledge or understanding about what will happen if the agent wishes to avoid blame. At most all that is required is (a) that the agent does believe or should believe that there is an alternative in which she avoids blame and (b) that if she were motivated to pursue that alternative she would in fact avoid blame in one way or another. This condition is satisfied in the Frankfurt cases. To take my jury example, Jones presumably believes that he can hold out and vote to acquit and that by doing so he would be fulfilling his duty, thus avoiding blameworthiness. If he

[132] Pereboom could argue that Fred is blameworthy for something a bit less specific, such as failing to rescue *a person*, rather than failing to rescue his daughter. I am not certain that our assessments of blame are so sophisticated. I find it intuitive to think that he is blameworthy under either description of his failure. He is also blameworthy for failing to try to rescue his wife, and perhaps for a number of other things. I do not think this is incompatible with him also being blameworthy for failing to rescue his daughter. In any case, I also do not think the possibility of a more generic description of his failure helps to save any sort of knowledge requirement for alternatives. In the case I am imagining, Fred does not have knowledge or awareness of an alternative in which he saves *a person*. He has a mistaken belief that there is an alternative in which he saves his wife. It is not as though, in the hypothetical situation, Fred could claim, "I knew that I would save *somebody* if I went in there." That might be the case if Fred had heard cries from within, or if someone had shouted, "There's someone trapped inside," and Fred had mistakenly inferred that it was probably his wife. But in the case I am imagining, Fred's false, specific belief about his wife does not imply a true, generic belief about someone or other.

were more strongly motivated to do his duty than he actually is, he would fail because of Black's intervention, but he would still succeed in avoiding blame. Thus, the alternative in these cases meets the relevant requirement.

Thus, although Fischer, McKenna, and Pereboom seek to identify a notion of robustness as a criterion for alternatives that are relevant to moral responsibility, their arguments cannot be used to show that the possibility of behaving differently, as identified by PAPB, is not necessary to moral responsibility. Given the affirmative argument for PAPB above, it would be quite surprising if they could have. Nonetheless, some of the claims made by Fischer and other defenders of a robustness requirement express a more general impression that the alternatives remaining in Frankfurt scenarios are of little relevance to moral responsibility. To address any lingering remnants of this feeling, I will close this section with a brief review of the positive significance of these ineliminable alternative scenarios.

First, the argument for PAPB connects the possibility of behaving differently to an important, and non-obvious, conceptual truth about moral responsibility. Such possibilities must exist not because of some accidental, yet-to-be-corrected flaw in the Frankfurt examples, but because their presence follows from the fact that one is responsible for behavior only when that behavior reveals that the strength of that agent's moral concern falls below or exceeds our reasonable expectations. Thus, far from being irrelevant, the presence of these alternatives tracks an important condition of moral responsibility for behavior.

Second, the principle is not trivial in its application. It can be used to identify a non-empty set of actions, omissions and other behaviors that are not grounds for moral praise or moral blame. First, non-voluntary behavior, behavior prompted by the direct manipulation of others, and actions that have no clear connection with moral considerations can, in general, be shown by PAPB to be behavior for which an agent is not responsible. If the strength of an agent's moral concern had nothing to do with the occurrence of a bit of non-voluntary behavior, then it is difficult to see how increasing or decreasing the strength of that moral concern would lead to relevantly different behavior. Second, actions that turn out badly, but which reveal only strong moral concern and innocently mistaken beliefs are also shown to be not deserving of blame. For if there has been no weakness in the agent's motivation to do what is right, then no reasonable improvement to the agent's moral motivation would produce a different result. Third, behavior prompted by compulsion or by a pathological desire is also ruled out as potentially responsible behavior. No reasonable change to the strength of the agent's moral motivation would produce a different result. PAPB and the truths underlying it help to provide us with a clear explanation of why these types of situations do not support responsibility.

Finally, PAPB helps validate and explain the usefulness of counterfactual reasoning as a tool in evaluating a person's responsibility. Some Frankfurtian compatibilists have suggested that Frankfurt cases show that how an agent would have behaved in other possible scenarios is not

relevant to moral responsibility. For example, Fischer has said,

> The Frankfurt-type examples . . . point us to something both remarkably pedestrian and extraordinarily important: moral responsibility for action depends on what actually happens. That is to say, moral responsibility for actions depends on the actual history of an action and not upon the existence or nature of alternative scenarios.[133]

It is not clear how far Fischer means to take this claim, as he has also expressly given an account of moral responsibility that depends on counterfactuals (just not the sort of counterfactuals suggested by PAP).[134] Fischer's suggested distinction between theories that focus on the actual history and those that focus on counterfactuals is a dubious one. But in any case, PAPB confirms that an effort to focus on the "actual history" without considering what else the agent might have done would be a mistake. Even if one could develop an accurate account of the conditions of moral responsibility that focuses entirely on what actually happens and does not rely on considerations about alternative possibilities, to abandon counterfactual tests would be to abandon the way that people actually reason when assessing responsibility. When we take a moment to reflect critically on a praising or blaming reaction, we naturally ask how easy or how difficult it would have been for the agent under consideration to have behaved otherwise. When we think about this question we imagine how much different

[133] Fischer (1994) p.158.

[134] See, for example, Fischer (1994) pp.166-67 (explaining the requirement of weak reasons-responsiveness in terms of alternate possible worlds); Fischer & Ravizza (1998) pp.63, 69-82 (explaining the requirements of weak and moderate reasons-responsiveness in terms alternate possible worlds and counterfactuals about the "mechanism" which produced the agent's actual behavior).

things would have to be in order for the person to have behaved differently. If it is clear that a more virtuous person in their position would also have done the same thing, we will (or at least we should) revise our initial instinct to blame.

While I have not tried to formulate PAPB as anything more than a necessary condition on responsibility, it is not difficult to see how the hypotheticals it has us consider can be tinkered with and modified to provide, if not a sufficient condition, a useful and important practical tool. PAPB applied as a necessary condition has us ask whether an improvement in the agent's desires would result in relevantly different behavior. It will not always be immediately obvious how much of a change it might be reasonable to expect. Sometimes we might be aware that a vast improvement to the person's will might have changed his behavior for the better, but we may be left with the sense that this is too much to expect. The fact that a person with a heroic work ethic might not have dawdled at all is not necessarily enough to show that the person deserves blame for dawdling just a little bit. Thus, if we start from this approach, the next natural question is what sorts of possible improvements to the will is it fair to expect of that particular person and measure him against. We might ask not just whether a person would have behaved differently if he were a hero, but whether he would have done so if he were more like other people of his age, his experience, his upbringing, and so on. Very often, we will ask, "Well would *I* have done any better in his situation?"

As PAPB and the argument for it indicate, it is no coincidence that such counterfactual reasoning can help guide our judgments about responsibility. If used properly, the effect of these questions is to show with increasing precision whether the agent's behavior reveals desires that we could reasonably expect him to have. Thus, while there is some truth to Fischer's claim that moral responsibility depends on what actually happens, PAPB reminds us that judgments about moral responsibility also depend on what might have happened instead. Therefore, I conclude that PAPB states a significant and interesting condition for moral responsibility.

III. "Ought" Implies "Can" and the Principle of Alternative Possibilities:

I want to close this Chapter by addressing an issue of internal consistency. In Chapter One, I argued for a particular version of the principle that "ought" implies "can." In this Chapter, I have suggested that the principle of alternate possibilities may be conceded to the Frankfurt examples, arguing only for PAPB instead. David Copp has argued that the principle of alternate possibilities, at least as it applies to blameworthiness, can be derived from "'ought' implies 'can.'" Thus, Copp's argument raises a possible issue of internal consistency. This section takes a close look at Copp's argument. I argue that one of its key premises is just as susceptible to the Frankfurt examples as PAP is. So, if Frankfurt examples do undermine PAP, they also undermine Copp's argument deriving PAP from "'ought' implies 'can'."

Copp's argument is fairly restated in step by step form as follows:

1) Suppose that a particular agent is morally

blameworthy for doing something.

2) If an agent is morally blameworthy for doing something, then it was wrong for her to have done that thing.

3) If it was wrong for an agent to have done something, then she ought to have done something other than what she did.

4) So, our particular agent ought to have done something other than what she actually did.

5) If an agent ought to have done something, then that agent could have done that thing.

6) So, our particular agent could have done something other than what she actually did.

7) So, if an agent is morally blameworthy for doing something, then that agent could have done something other than what she actually did.[135]

This statement of the argument makes each of the steps explicit, but the idea can be summed up more simply. In addition to the idea that "ought" implies "can," the argument really requires only one assumption: the idea that if a person is blameworthy for doing something then she ought to have done something else. In Copp's argument, this step is presented as premises 2 and 3, mediated by an additional step through the concept of wrongfulness. We should keep that in mind, but by leapfrogging over that step we can see the simplicity and appeal of the argument in the following brief summary: If a person is blameworthy for doing something, then she ought to have done something else, and if she ought to have done something else, then she could have done something else. So, if she was blameworthy for doing something,

[135] See Copp (2003) p.270 and Copp (1997) p.445.

then she could have done something else.

Copp's argument makes no mention of which senses of "can" and "able" are relevant, but of course, the version of "'ought' implies 'can'" that I have defended involves a very specific sense of "can." In Section VII of Chapter One, I identified three distinct factors that make it possible for an agent to perform an action: whether the agent has the ability, whether she is in a situation constituting an opportunity to exercise that ability; and whether she is sufficiently motivated to exercise the ability. Given these distinct factors, I argued "ought to ϕ" implies "can ϕ" only in the sense that ϕ-ing is compatible with the agent's abilities and opportunities, but not necessarily with her motivation. This fact affects the conclusion of the argument, but it does not appear to affect the validity. We can make the proper understanding of "'ought' implies 'can'" explicit in Copp's argument with a simple revision to the last three steps of the argument:

> 5*) If an agent ought to have done something, then that agent could have done that thing, in the sense that she had the ability and the opportunity to do that thing.
>
> 6*) So, our particular agent could have done something other than what she actually did, in the sense that she had the ability and the opportunity to do so.
>
> 7*) So, if an agent is morally blameworthy for doing something, then that agent could have done something other than what she actually did, in the sense that she had the ability and opportunity to do so.

It is fair to call this conclusion a version of the principle of alternate

possibilities, at least for blameworthiness.

Without more controversial assumptions, Copp's argument cannot be extended to support PAP for praiseworthiness. If a person is praiseworthy for doing something, then she presumably did what she ought to have done. So, the fact that "ought" implies "can" cannot be used to show that she could have done anything else. Instead, we would need something like the principle that "ought" implies "can do otherwise." Some philosophers do accept this principle.[136] But it is neither as intuitively compelling nor as widely accepted as "'ought' implies 'can'." I have not argued for it at any point. So, I shall focus on the limited version of PAP supported directly by Copp's argument, a version limited to blameworthiness.

Is there a problem with accepting this modified version of Copp's argument? I have argued in favor of an alternative to PAP, but I have not, thus far, given PAP itself much of a chance. One might suppose that the version of PAP derived from my version of "'ought' implies 'can'" can survive the Frankfurt examples. The distinctions between abilities, opportunities, and motivation help to distinguish three kinds of causally relevant factors. Abilities are powers of a sort, generally grounded in certain properties of the agent, though not only intrinsic properties. They are also commonly grounded in the role a particular agent plays in a social structure or in things that agent

[136] See Haji (1999). Hare (1963) also provides one possible basis for thinking that "ought" implies "can do otherwise." Hare argues that the question of what an agent "ought" to do only arises when there is what he calls a "practical issue," a question of what to do. One could argue that there is only an *issue* about how to act if there is more than one way of acting open to the agent. The action the agent "ought" to perform is simply the most favorable of the ones that are open to him, implying that there is some other way of acting that is open to him.

possesses. Opportunities have more to do with the agent's circumstances and environment. They exist when the conditions are appropriate for the exercise of an ability. Motivation, obviously, has to do with the agent's desires. It determines whether or not the agent will attempt to exercise an ability. So divided, it could be argued that the Frankfurt examples assure that an agent will not do otherwise only by ensuring that he will not be adequately motivated to do otherwise. By intervening, in the counterfactual situation, in the neurological and psychological processes, Frankfurt's interveners ensure that Jones will not be motivated enough to exercise an ability to do anything else. So, one might suppose that the Frankfurt cases are cases in which responsible agents are unable to do otherwise only in the sense that they cannot become adequately motivated to do otherwise. They do not, on this view, interfere with the agent's ability or opportunity to do otherwise.

I do not think this suggestion can be sustained, at least not without a more restrictive account of abilities and opportunities than the one I favor. While it is true that Frankfurtian interveners typically prevent agents from becoming sufficiently motivated, it is also very intuitive to suppose that they prevent the agent from having an opportunity to do otherwise. It seems odd to say that an agent has a opportunity to φ when a super-powered meddler stands by, ready, willing, and able to do anything necessary to prevent the agent from φ-ing, even if the meddler's plan is to interfere primarily with one's motivation. In addition, there is nothing about my account of abilities and opportunities that suggests otherwise. What constitutes an opportunity to φ on

my account depends on the fully specified content of the ability to φ, and the fully specified content of an ability is virtually unlimited. It can contain countless conditions that we have never had any reason to notice or consider before. So, if it is natural to say that one lacks the opportunity to φ when a Frankfurtian intervener is prepared to prevent one from φ-ing, then we have every reason to suppose that the absence of a Frankfurtian intervener is a condition included in the fully specified content of the ability to φ.

Assuming then that the conclusion of Copp's argument is false, how do we avoid the inference that "'ought' implies 'can'" is also false? In the short form version of the argument I offered above, there is only one premise other than the principle that "ought" implies "can": the claim that if an agent is blameworthy for doing φ then she ought to have done not-φ. This premise is as susceptible to Frankfurt cases as the conclusion, PAP. Frankfurt cases have made it implausible to suggest that if a person is blameworthy for doing something, then she could have *done* otherwise. For just the same reasons, they make it implausible to suggest that if a person is blameworthy for doing something, then she ought to have *done* otherwise. All that is required for blameworthiness is that the agent would have *behaved* differently if her moral motivation was stronger. There is no reason to infer from blameworthiness anything more than that the agent ought to have been better motivated. Under the analysis in Chapter One, this is not the sort of "ought to do" claims that imply "can." A claim about how one ought to *be* motivated does not entail any particular claim about what one ought to do.

Recall that, as Copp presents the argument, the inference from blameworthiness to "ought to have done otherwise" takes two steps. First, he claims that if one is blameworthy for doing something, then it was wrong for the agent to have done it, and second, he claims that if it was wrong for an agent to have done something, then he ought to have done something else. It is the second of these two specific steps that is implausible.[137] Recall Jones's decision to vote to convict Red. It was wrong for Jones, out of reluctance to stand up to his peers, to vote to convict Red without adequate evidence of guilt. But what else was Jones supposed to do? Surely he should have been more principled or more concerned to do the right thing, but if he were, Black would have ensured that Jones did nothing at all. So, from the fact that a person acted wrongly, we cannot safely infer that she should have *done* something else. From the fact that it was wrong of a person to do φ, the only safe inference is that she should have been better motivated, *i.e.*, more

[137] Gideon Yaffe has objected to this idea as well, similarly pointing out that in some circumstances a person might discharge an obligation not to X by doing nothing at all, e.g., when he is knocked unconscious just before the opportunity to X occurs. Yaffe (1999). Yaffe's approach (and by implication, my own approach) is criticized in Fischer (2003). Fischer's rejection of Yaffe's approach turns on an assertion that if one ought not to X, then one must have "genuine access" to a possible world in which one does not X, where "genuine access" means that the possible world has the same laws and history as the actual world. Fischer appears to believe that "ought implies can" can be squared with common sense moral claims about what a person ought to do only if indeterminism is true. Granting his assumption, we could only say that A should not have tortured B as he did if there was a possible world with the very same laws and history in which A did not torture B. This would require indeterminism, or perhaps, miracles. Fischer offers no discernible argument for this very bold and controversial assumption, except to repeatedly say that it "seems to [him]" that anyone accepting "ought implies can" should believe this and that it somehow follows from the "motivation" behind "ought implies can." Fischer does not explain what he sees as the "motivation" behind the principle or why it might support his bold assertion. Fischer concludes by rejecting "ought implies can" – not surprisingly, given his assumptions about the principle. I have defended a version of "ought implies can" and explained the motivation behind it. It provides no support for Fischer's assumption.

disposed to do her duty. But being better motivated guarantees no more than that she will not voluntarily do what she actually did. It does not ensure that she has the opportunity to *do* anything different, and so we should not automatically assume that she should do anything different.

I do not dispute that if a person is blameworthy for doing something or if it was wrong of her to do it, then we can and should expect her to have behaved better. But Copp's argument assumes that this means that the agent should have *done* something else. In most cases, this is a fairly safe assumption. In most cases, we may safely assume that there is no counterfactual intervener waiting in the wings, and so all that would be needed for her to do something else is for her to have cared enough to do it. But there are exceptional cases, and these exceptional cases show the limits of what blameworthiness requires. A person can be blameworthy for what she did even though she could not have *done* otherwise. In these cases, the intuition that we may demand or expect better of the agent is satisfied by the fact that her will should have been better and by the fact that had her will been better, her behavior would have been different. Therefore, Copp's argument does not show any inconsistency between my defense of "'ought' implies 'can'" and my rejection of PAP in favor of PAPB.

IV. Epilogue: Determinism and Reasonable Expectations for Moral Motivation

In explaining and defending PAPB, I have relied in various places on a notion of reasonable expectations, but thus far, I have avoided addressing questions about this idea. As I noted at the end of Chapter Two, the idea that

a person might be reasonably expected to have stronger (or weaker) moral motivation might be problematic, particularly in a deterministic world. I cannot answer all of the questions one might ask about reasonable expectations, but I want to say a bit about their role in PAPB and then at least outline the problems they raise.

PAPB incorporates the idea that it is reasonable to expect a person to be differently motivated in some respects, that it is reasonable to demand that she be more concerned about reasons or to be unsurprised if she had been less concerned. I have invoked reasonableness as a limitation on what we can expect of an agent in terms of moral motivation, and this limitation is crucial to constraining the range of cases in which we will find a person to be blameworthy or praiseworthy. In the case of blameworthiness, the relevant reasonable expectations are normative. They specify the amount of moral motivation that we believe a particular person ought to have, but they are reasonable in so far as they take appropriate account of human capacities for moral motivation. In a particular case, it might be that a person would have behaved differently if she had a superhuman commitment to moral reasons, but it is not reasonable to say she deserves blame because she lacks this superhuman commitment. Nor will we suppose that a person deserves blame if he is too immature, insane, or mentally deficient. It is unreasonable to expect a person who cannot even grasp moral concepts to be concerned and motivated by moral reasons. And it can also be unreasonable to expect as much moral concern from those who have not received an adequate moral

education as we can expect from those who have. Though PAPB would still state a necessary condition for responsibility without this reasonableness limitation, it is a more interesting and informative principle with it. It does a better job of tracking blameworthiness.

In the case of praiseworthiness, reasonable expectations play a limiting role in a different way. A praiseworthy agent is one who has met or exceeded our reasonable normative expectations for moral motivation. But I have suggested that it also must be reasonable, in some other sense, to expect worse of the agent. It is certainly not the case that we might have a normative expectation of less moral motivation, and it need not be the case that we have grounds for an actual, all things considered, prediction for less. Instead, the point is that it would be reasonable to expect less, in a predictive sense, given our experience with the temptations to which she and other human beings are normally susceptible. On the basis of this information, we may not literally believe that she will succumb to these temptations, but we believe that it would be understandable (either for her or for people in general) if she did. Only if we can see that behaving otherwise would satisfy some other desires of hers, or some other desires that she would be likely to have if she were not such a good person, will we suppose that the behavior displays the sort of strong moral concern that merits praise. This limits the cases in which a person will be praiseworthy, as it means that a person will not merit praise for avoiding the sorts of bad acts that neither she nor any other normal person would have ever been tempted to perform in the first place.

But one might ask why it is ever reasonable to expect a person to have different moral motivation at all. Determinism can make this question seem particularly pressing. (This is not to say that the question would be any easier to answer if indeterminism were true.) If determinism is true, then a person can be no more and no less concerned than the laws of nature and the past entail. She is always as concerned about doing the right thing as she possibly can be. What makes it reasonable to expect any more? One might also ask what makes moral motivation different than the other kinds of changes that we cannot reasonably demand of a person? I have argued above that we take abilities, opportunities, and impulses to be, to a certain extent, given. We do not expect or demand that a person has different natural talents than those that are given to her in the natural lottery, so to speak. We might expect her to develop those talents, but only to the extent that she could, if she cared to. She cannot be expected to develop talents that exceed her natural potential. So, why is moral motivation not also treated as part of the natural lottery, something that is simply given to her and that we must simply accept and work with? If determinism is true, one's potential for moral motivation is no less determined than anything else. Why can we expect more than what an agent has, given the laws of nature and history?

We could note that our reasonable expectations of moral motivation are "ought to be" claims: a blameworthy person ought to be more concerned to do the right thing. As "ought to be" claims they do not imply that the person *can* be more concerned, at least not according to the principle discussed in

Chapter One. But this point is purely negative. It rules out an argument against the appropriateness of these expectations, but it does not affirmatively show why it is reasonable to have them. Moreover, while these expectations may not trigger the "ought" implies "can" principle, they do reflect a sensitivity to what is possible or realistic. As I have just pointed out, it is unreasonable to demand superhuman commitment or to expect moral motivation from a person with no capacity for understanding moral concepts. So why are they not also sensitive to the limits that determinism imposes?

We might also point out that it seems essential to moral realism to believe that people should be concerned to do the right thing. To see the world in moral terms is to believe that people ought to be concerned and motivated to do the right thing and promote valuable states of affairs. To accept that there is good and bad, right and wrong, is to accept that people ought to be concerned about these facts. Thus, the reasonableness of expecting moral concern and moral motivation is, arguably at least, a stipulation of the moral world view. Therefore, to question the propriety of expecting moral concern may simply be to ask whether it makes sense to see the world in moral terms. And that may not be a question that we can answer based on moral principles about whether people do or do not deserve to be held responsible. But I do not find this response to be altogether satisfying either. Without questioning whether expectations of greater moral concern ever are reasonable, one could reasonably demand some account of when and why they are. This should be part of a full account of moral responsibility.

Further, it is one thing to believe that people ought to be more motivated to do the right thing. It is another to demand that they be more motivated and to hold them accountable if they are not. We may still ask what makes this demand, backed by blame, reasonable. That is not to ask whether it makes sense to see the world in moral terms, but only whether it makes sense to hold people responsible.

So, if we accept PAPB as a necessary condition for responsibility, we might still demand an explanation of when and why it is reasonable to expect a person to have any particular degree of moral motivation other than what they actually have. For my own part, I am inclined to believe that if a person is a competent, mature agent who has had some basic moral education (providing a basic appreciation of the difference between right and wrong, fostering a feeling of concern and sympathy for others, *etc.*), it is presumptively fair and, therefore, reasonable to demand a good degree of moral motivation. I see no clear reason for supposing that this should not be so, but I also cannot now offer a positive account of why it should be so. My purpose in this Chapter has been to argue for a necessary condition for moral responsibility that links responsibility with alternate possibilities without entailing that determinism and responsibility are incompatible. My purpose has not been to address every concern about moral responsibility and determinism, or to provide a positive explanation for why the two are compatible. Having done so, and thus having defended a position on which the argument from PAP to incompatibilism fails, it seems fair to conclude that

we should not accept incompatibilism about determinism and responsibility unless there is some other argument for that conclusion. At the same time, it is also fair to ask for a more satisfying positive explanation of why determinism and responsibility are compatible, but that must remain work for another occasion.

REFERENCES

Anselm of Canterbury (1988): "On Truth", in *The Major Works*, edited by Brian Davies and Gillian Evans, Oxford: Oxford University Press.

Arpaly, Nomy (2006): *Merit, Meaning and Human Bondage: An Essay on Free Will*, Princeton, New Jersey: Princeton University Press.

Austin, J.L. (1956): "Ifs and Cans", *Proceedings of the British Academy*, Vol. 42, pp.109-32.

Ayer, A.J. (1954): "Freedom and Necessity", in *Philosophical Essays*, London: MacMillan.

Broad, C.D. (1952): "Determinism, Indeterminism, and Libertarianism" in *Ethics and the History of Philosophy: Selected Essays*, London: Routledge & Kegan Paul.

Copp, David (2003): "'Ought' Implies 'Can', Blameworthiness, and the Principle of Alternate Possibilities", in *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (Widerker & McKenna Eds.), Burlington, Vermont: Ashgate Publishing Co.

Copp, David (1997): "Defending the Principle of Alternate Possibilities: Blameworthiness and Moral Responsibility", *Nous*, Vol. 31, pp.441-456.

Davidson, Donald (1971): "Agency", in *Essays on Actions and Events* (1980), Oxford: Oxford University Press.

Davidson, Donald (1967): "The Logical Form of Action Sentences," in *Essays on Actions and Events* (1980), Oxford: Oxford University Press.

Fischer, John Martin (2003): "'Ought-Implies-Can,' Causal Determinism, and Moral Responsibility", *Analysis*, Vol. 63, pp.244-250.

Fischer, John Martin (1994): *The Metaphysics of Free Will*, Cambridge, Mass.: Blackwell Publishers.

Fischer, John Martin & Ravizza, Mark (1998): *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

Frankfurt, Harry G. (1971): "Freedom of the Will and the Concept of a Person," *Journal of Philosophy*, Vol. 68, pp.5-20.

Frankfurt, Harry G. (1969): “Alternate Possibilities and Moral Responsibility,” *Journal of Philosophy*, Vol. 66, pp.829-839.

Graham, Peter (2011): “‘Ought’ and Ability,” *Philosophical Review*, Vol. 120, pp. 337-382.

Gert, Bernard & Duggan, Timothy (1979): “Free Will as the Ability to Will”, *Nous*, Vol. 13, pp.197-217.

Gert, Bernard & Duggan, Timothy (1967): “Voluntary Abilities”, *American Philosophical Quarterly*, Vol. 4, pp.127-135.

Haji, Ishtiyaque (1999): “Moral Anchors and Control,” *Canadian Journal of Philosophy*, Vol. 29, pp.175-203.

Harman, Gilbert (1977): *The Nature of Morality*, New York: Oxford University Press.

Hume, David (1740): *An Enquiry Concerning Human Understanding*, Edited by Eric Steinberg (2d Ed. 1993), Indianapolis: Hackett Publishing Co.

Kane, Robert (1996): *The Significance of Free Will*, New York: Oxford University Press.

Kant, Immanuel (1781/1787): *Critique of Pure Reason*, translated by Paul Guyer & Allen Wood (1997), Cambridge: Cambridge University Press.

Lewis, David (1976): “The Paradoxes of Time Travel”, *American Philosophical Quarterly*, Vol. 13, pp. 145-152.

McKenna, Michael (2003): “Robustness, Control, and the Demand for Morally Significant Alternatives: Frankfurt Examples with Oodles and Oodles of Alternatives”, in *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (Widerker & McKenna Eds.), Burlington, Vermont: Ashgate Publishing Co.

Mele, Alfred R. (2006): *Free Will and Luck*, New York: Oxford University Press.

Moore, G.E. (1922): “The Nature of Moral Philosophy” in G.E. Moore, *Philosophical Studies*, New York: Harcourt Brace.

Moore, G.E. (1912): *Ethics*, London: Oxford University Press.

O’Connor, Timothy (2000): *Persons and Causes*, Oxford: Oxford University Press.

Parfit, Derek (1984): *Reasons and Persons*, Oxford: Oxford University Press.

Pereboom, Derk (2005): "Defending Hard Incompatibilism", *Midwest Studies in Philosophy*, Vol. 29, pp.228-247.

Pereboom, Derk (2001): *Living Without Free Will*, Cambridge: Cambridge University Press.

Pereboom, Derk (1997): "Determinism Al Dente" in *Free Will* (2d Ed. 2009), Indianapolis: Hackett Publishing Co.

Prior, Elizabeth (1985): *Dispositions*, Aberdeen: Aberdeen University Press.

Saka, Paul (2000): "*Ought* Does Not Imply *Can*," *American Philosophical Quarterly*, Vol. 37, pp.93-105.

Scanlon, T.M. (2008), *Moral Dimensions: Permissibility, Meaning, Blame* Cambridge, Mass: Harvard University Press.

Scanlon, T.M. (1998): *What We Owe To Each Other*, Cambridge, Mass: Harvard University Press.

Sher, George (2006): *In Praise of Blame*, New York: Oxford University Press.

Sinnott-Armstrong, Walter (1984): "'Ought' Conversationally Implies 'Can'", *The Philosophical Review*, Vol. 93, pp.249-261.

Strawson, Galen (1994): "The Impossibility of Moral Responsibility", Philosophical Studies, Vol. 75, pp. 5-24, cited as reprinted in Free Will (Gary Watson (Ed.), 2d Ed. 2003) Oxford: Oxford University Press.

Strawson, P.F. (1962): "Freedom and Resentment", Proceedings of the British Academy, Vol. 48, pp.1-25, cited as reprinted in Free Will (Gary Watson (Ed.), 2d Ed. 2003) Oxford: Oxford University Press.

Streumer, Bart (2007): "Reasons and Impossibility," *Philosophical Studies*, Vol. 136, pp.351-384.

Van Inwagen, Peter (1983): *An Essay on Free Will*, Oxford: Oxford University Press.

Vihvelin, Kadri (2000): "Libertarian Compatibilism", *Philosophical Perspectives*, Vol. 14, pp.139-166.

Vranas, Peter (2007): "I Ought, Therefore I Can," *Philosophical Studies*, Vol.137, pp.167-216.

Yaffe, Gideon (1999): "'Ought Implies Can and the Principle of Alternate Possibilities," *Analysis*, Vol. 59, pp.218-22.

Zimmerman, Michael (1996): *The Concept of Moral Obligation*, New York: Cambridge University Press.