RESPONSIBILITY, IDENTITY, AND LUCK

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

Patrick Timothy Mayer

January 2013

© 2013 Patrick Timothy Mayer

RESPONSIBILITY, IDENTITY, AND LUCK

Patrick Timothy Mayer, Ph. D.

Cornell University 2013

In this dissertation I argue against some popular arguments for the compatibilism of moral responsibility and determinism, and for a compatibilist account of moral responsibility and agency of my own. In the first chapter I argue against Strawsonian inspired accounts of moral responsibility that would, if accurate, provide reason to reject incompatibilism about moral responsibility and determinism. In the second chapter I argue that Quality of Will accounts of moral responsibility, while likely correct, do not provide reason to rejection incompatibilism. In the third chapter I provide and argument for incomaptibilism that relies on none of the controversial assumptions about control and the ability to do otherwise which have presented problems for arguments for incompatibilism in the past. Thus, this argument is immune to the criticisms of traditionally incompatibilist accounts of control and ability. In the last chapter I present a compatibilist account of moral agency that I endorse. This accoutn is a combination of two popular accounts of moral agency that I show fail, when taken on their own.

BIOGRAPHICAL SKETCH

Patrick Timothy Mayer received his Bachelor of Arts degree in Philosophy and Political Science from the University of Maryland, Baltimore County before doing his graduate study at the Sage School of Philosophy of Cornell University.

To Leah, whose sacrifices made this possible,

Olivia, whose arrival made this worth doing,

And Robin, Ralph, and Eric for their love, support and guidance.

ACKNOWLEDGMENTS

The work included here is the descendant of work I began as an undergraduate and so I would like to thank the faculty of the Philosophy Department of the University of Maryland, Baltimore County for their years of support. I would especially like to thank Dr. Stephen Yalowitz, who first introduced me to the problems surrounding free will and moral responsibility and supervised my undergraduate thesis on the subject. I want to express my gratitude to the entire faculty and graduate student body of the Sage School of Philosophy for years of helpful and enjoyable conversation on the wide range of topics that found their way into this dissertation. Deserving of special thanks are Dr. Michael Fara, Dr. Terence Irwin, Dr. Scott MacDonald, Dr. Michelle Kosch and Dr. Derk Pereboom, who supervised the dissertation, reading hundreds of pages of drafts and providing insightful and challenging commentary throughout, even when the drafts themselves might not have merited such close attention.

TABLE OF CONTENTS

Introduction: Responsibility, Luck and Identity

This dissertation is about moral responsibility. It aims to clarify the issues that separate compatibilists about moral responsibility and determinism, and incompatibilists about moral responsibility and determinism. My hope is that such clarification will reveal the ways in which compatibilists can go about improving their arguments and views. I have been a compatibilist for almost as long as I have known about the compatibility problems that face moral responsibility and free will. As such I have always been very sensitive to claims to the effect that compatibilists have nothing to offer but shallow accounts of responsibility and freedom. I feel that these attacks applied generally to compatibilism are inapt, but I understand why people are tempted to make them. Classical compatibilists, by which I mean compatibilists before the 1960's, tended to offer defenses of their view that either completely failed to show any significant understanding of the central concepts or seemed nothing better than desperate logical or semantic ploys to avoid having to deal with the real issue. So some compatibilists seemed to think that freedom consisted entirely in getting what you wanted[1] and that a person's being responsible for an action consisted entirely in it being good social policy to punish or reward them for it.[2] These accounts are either clearly wrong, or accounts of concepts that are not at issue in the free will debate. For the purposes of this dissertation I will say that it is the former option that is correct.[3]

There are many compatibilists who avoid giving false accounts of the central

[1] Hobbes (1994)
[2] Many compatibilists, notably Hobart(1934), Moore(1993), Nowell-Smith (1948) etc.
[3] While I do not mind lush ontologies I am allergic to a set of concepts that is any larger than it absolutely has to be. It is possible, by appeal to a permissive principle of concept production, to avoid ever giving an incorrect account of something, by dealing with any purported counter-example to the account as actually being about some other concept easily conflated with the concept you are providing an account for. I think that this tendency to immunize our work from counter-examples just leads to confusions, so as far as is possible I will restrict myself to talking about a shared concept of moral responsibility, not different concepts had by compatibilists and incompatibilists.

concepts. Many of them do so by not giving any. One way to argue for compatibilism is simply to argue that the central incompatibilist arguments are unsound. Traditional incompatibilist arguments like the Consequence Argument rely on premises and inference principles that are controversial.[4] That is hardly surprising since their conclusions are so startling. These premises and inference principles have been subjected to sustained critique and in many cases they have been shown to fail. What is more important than the fact that they fail, however, is the way they fail. In short, they fail for reasons that do not seem like evidence for the claim that determinism is compatible with freedom or responsibility. So we have Lewisian compatibilism which started out as little more than the argument that one popular incompatibilist argument, the Consequence Argument, is invalid.[5] The Consequence Argument has since been revised so that it is valid, and there are other traditional incompatibilist arguments that Lewis' argument never had a chance to be successful against. Since that time Lewisian compatibilists have tried to develop an account of free action based on Lewis' sparse comments on the issue. This account of free action depends on an account of ability that incompatibilists have long since made clear is not the one they accept, and which is subject to significant counter-examples, the conditional account of ability.

That the conditional account of ability is denied by incompatibilists, and that some popular versions of it face significant counter-examples does not by itself mean that it should be abandoned, but it is worth noting that it provides the entirety of the Lewisian account of free action. The fault that Lewis identified in the Consequence Argument does not suggest that we are free, it simply says that the Consequence Argument equivocates on 'ability'. There are clearer examples of compatibilist

[4] The earliest version of the Consequence Argument that I am aware of is in Ginet (1966), though he does not identify it by that name. The most famous presentation of the Consequence Argument is in Van Inwagen (1983)

[5] Lewis (1981)

responses to incompatibilist arguments that don't really seem like they do anything to give evidence for compatibilism except knock down incompatibilist arguments. And if one assumes that the burden of proof rests with the incompatibilist, that is a perfectly valid strategy. But to rest on the fact that one does not have the burden of proof, without any further explication of one's position is to simply take it for granted that common sense is probably getting things right. And this is what might seem shallow. After all we don't even know what the common sense view of freedom and responsibility is, so how can we have reasonable confidence in it just because no one has managed to prove beyond a shadow of a doubt that it is confused or inapplicable? Such a reliance on the common sense notion produces in some incompatibilists the sense that compatibilists are not truly engaging the issues that are raised by their arguments.

This is not to suggest that we ought to treat unsound arguments as though they were sound. We shouldn't, and to the extent that compatibilists have undermined particular incompatibilist arguments we should abandon those incompatibilist arguments. But the proper response to, for example, Lewis showing that an early version of the Consequence Argument was unsound, is not to abandon the Consequence Argument. Lewis gave no reasons to think that any future Consequence Argument would have to be unsound for similar reasons. Similarly counterexamples to the transfer of powerlessness principle do not constitute counter-examples to related inference principles, such as the transfer of lack of responsibility principle. The essential criticism here, which applies to much but not all compatibilist responses to incompatibilist arguments, is that they do not show a concern for the central point that incompatibilists are trying to express with their arguments. Continually finding ways to poke holes in incompatibilist arguments is not enough. What is needed is a compatibilist response which gives compelling reasons to think that all future

incompatibilist arguments will also be unsound.

In the first part of this dissertation I will look at some compatibilist arguments that aim to do just that. The general strategy that will be examined is moving from an account of praiseworthiness and blameworthiness to an account of moral responsibility on which compatibilism must be true. I will argue that none of the instances of this strategy that have been presented succeed. They either present faulty accounts of praiseworthiness and blameworthiness, or the accounts they offer do not imply that compatibilism must be true. The more general problem is that these Strawsonian arguments fail to properly appreciate the nature of incompatibilism and incompatibilist arguments. If the earlier compatibilist strategies were guilty of focusing too much on responding to particular arguments, the Strawsonian strategy is guilty of focusing too little on those arguments. It is not clear what a Strawsonian could say in response to the Consequence Argument. Perhaps she could say that whatever the merits of the argument it must be unsound because of the Strawsonian argument for the conclusion's falsity. Or more likely she might say that to make theoretical arguments about the issue of moral responsibility is to commit some kind of category mistake. Neither response is compelling and neither takes seriously, in the least, the point that the incompatibilist is trying to express, as I will show.

In the second part of the dissertation I address the more piecemeal responses to incompatibilism, and I do so by sidestepping them completely. These piecemeal responses typically involve discussion of when and why people have the ability to do otherwise or control over their actions. There are compatibilist-friendly accounts of ability and control which imply that determinism is compatible with the ability to do otherwise and control over one's actions. There are also well known compatibilist arguments against the claims that the ability to do otherwise and control over one's actions are necessary for moral responsibility. These arguments and accounts have

generated massive amounts of literature and I could not hope to sort through it all in one dissertation. With that in mind I offer an argument for incompatibilism that does not rely on any claims about ability or control. If the argument is valid it shows that the incompatibilist can grant the compatibilist all her controversial claims about ability and control without granting anything that precludes her giving a sound argument for incompatibilism. The Moral Luck Argument thus marks a significant dialectical advance over previous incompatibilist arguments.

In the first part of the paper I deal with overly general defenses of compatibilism. In the second I deal with overly particular defenses of compatibilism. In the third section I deal with those defenses of compatibilism that I think are of the right kind. These defenses consist of an appeal to a general account of what it is for an agent to be morally responsible for an action, just as the Strawsonian strategy did, but the accounts they appeal to allow them to respond in a focused way to incompatibilist arguments, as the piecemeal compatibilist responses do. The two defenses I examine I will, following others, call the Real Self View and the Reasons Responsiveness view.[6] I argue that both views, while being the right kind of view to offer in a defense of compatibilism, fail. Each view faces extensional and explanatory problems and it is not clear how, on each view's own resources those problems could be resolved. Despite these problems, however, I argue that something in the ballpark of these two views gives compatibilists the best chance to defend their view. In short, while neither view can save itself, each view has something to offer the other that bolsters it. I will argue that a compatibilist account of moral responsibility and free action which combined the best elements of the both the Real Self view and the Reasons Responsiveness view is plausible. What is more I will sketch the ways that such an

[6] 'The Real Self View' is a name provided by Susan Wolf (1993) for a view which was first put forward in the contemporary literature by Frankfurt (1971). The Reasons-Responsiveness View was first put forward under that name by Fischer and Ravizza (1998)

argument can present responses to incompatibilist arguments, both the traditional ones and my own, that are dialectically effective and logically sound.

So in essence I aim to show how to be and how not to be a compatibilist. I do not intend to establish that compatibilism is true, just give reasons why certain strategies for establishing it should be pursued and others abandoned. Along the way I argue for much else. The secondary goal of this dissertation is to show the ways that a proper resolution of the debate about moral responsibility depends on substantive issues in ethical theory, meta-ethics and moral psychology.

This is not terribly surprising. The problem of moral responsibility has always been about how it is appropriate to treat people given their behavior, and the problem of free will has always been linked with the problem of moral responsibility in such a way that inclines theorists to think that what makes someone free is in large part what makes her morally responsible for her actions. So any proposed account of what makes someone morally responsible that did not have ethical implications would for that reason be an inadequate view. And it is also very clear that issues in moral psychology and meta-ethics entail and are entailed by certain principles in ethics. Some of these implications are well known. It is, for example, commonly accepted that certain accounts of how we deliberate are inconsistent with libertarian accounts of agency. If we act according to our strongest desire, and our strongest desires were always unconscious desires, or hardwired desires, then libertarianism would be false. But of course few people have seriously defended such a view. It is commonly acknowledged that if one denies the possibility of desert in general, perhaps because one accepts some very strong version of consequentialism, then one must accept some version or other of compatibilism. Certainly some kinds of anti-realism about the normative or evaluative would force us to abandon saying things like 'If determinism obtains then no one is blameworthy or praiseworthy for anything.'

I think the linkages are far more common than this, and they do not consist just of the entailments of extremely implausible views like the ones I just mentioned. In this dissertation I will show how debates in axiology matter for giving a proper account of moral responsibility. I show how issues about what is and can be fair, matter for whether anyone actually is responsible for her actions. And the disputes I bring up are live disputes. For example, I claim that on the best account of moral responsibility available to us, hedonism about the good must false. I also appeal to contentious moral claims in the process of presenting a new argument for incompatibilism. I will not, in this dissertation, present arguments for these contentious moral principles that I think ought to convince ethical theorists convinced of their falsity. This is not a dissertation on ethics. What it aims to do is show the argumentative burden that compatibilists are under is significant and show that part of that argumentative burden includes arguing for contentious ethical principles.

Chapter 1: Praise and Blame

We are all familiar with arguments purporting to show that determinism is incompatible with moral responsibility. We are also familiar with arguments that would, if sound, show that determinism is incompatible with moral responsibility, but only when supplemented with further very plausible premises. In the first group of arguments are Van Inwagen's Direct Argument[7], Galen Strawson's Impossibility Argument[8], or Pereboom's Four Case Argument[9]. In the second group of arguments are Van Inwagen's Consequence Argument[10], and Haji's Deontic Argument.[11][12] Now I think that all these arguments are unsound, but I think that they are all unsound for different reasons. Each argument advances some particular claim about moral responsibility, or freedom of the will, or the nature of moral obligation, that I think is false. I don't think there is some general argument for the compatibility of moral responsibility and determinism that can be used effectively against all these incompatibilist arguments.[13]

Not all compatibilists agree. Some compatibilists influenced by Peter Strawson[14], and probably Strawson himself[15], think that if one really understands the

[7] Van Inwagen (1980)
[8] Strawson (1994)
[9] Pereboom (2001)
[10] Van Inwagen (1983)
[11] Haji (2002)
[12] The Consequence Argument need only be supplemented with the premise that freedom of the will is a necessary condition of moral responsibility, while the Deontic Argument need only be supplemented with the premise that for a person to count as morally responsible for an action that action must meet or fail to meet the demands of some moral obligation.
[13] At least not one that is not question begging. If a compatibilist asserted a particular account of free action she could show, from the assumption of that account, that freedom and determinism are compatible, and with a little effort she could also show that moral responsibility and determinism are compatible. But the incompatibilist will simply deny the proposed account of free action. In denying it the incompatibilist will point to the very principles she appeals to in making the standard incompatibilist arguments. So to defend the compatibilist account of free action it would be necessary to say why the controversial premises in the incompatibilist arguments are false.
[14] Wallace (1994), MaGill(1997), Watson (1987), Scanlon (2008), McKenna (1998) and (2005), Arpaly (2006) are the authors that I discuss here, because they are explicit in the debt they owe to Strawson's work. There are very likely more Strawsonian compatibilists than this. Many for instance cite

nature of moral responsibility, one will see that determinism's obtaining simply can't give us reason to stop holding people morally responsible. As I have said I do not think there is any such general argument for compatibilism. I believe the best case for compatibilism is a piecemeal one. It consists of good arguments against the key premises of the arguments mentioned above, along with any other significant incompatibilist arguments.[16] The Strawsonian arguments, however, seek to do away with incompatibilism without telling us what is wrong with standard incompatibilist arguments.[17] Strawsonians are committed to thinking there is some kind of confusion about the nature of moral responsibility on the part of incompatibilists which explains why they offer the kinds of arguments they do, and so that the elimination of this confusion about what we are talking about is sufficient to show that the arguments are no good, or pointless, or something like that.

In this paper I will examine three Strawsonian accounts of moral responsibility, the Emotion Account, the Relationships Account, and the Quality of Will Account. I argue that these Strawsonian accounts of moral responsibility are either false, or do not show what Strawsonians take them to show about the coherence and tenability of incompatibilism. The goal to which this paper is in service is showing that the correct account of what moral responsibility is will be neutral between compatibilist and incompatibilist positions about moral responsibility, determinism and indeterminism. While the arguments in this chapter do not prove that

---

approvingly the relationship (whatever it is) that Strawson posited between moral responsibility and the reactive attitudes. Fewer authors do much more with that connection than cite it however.

[15] Strawson (1974)

[16] This sounds as though I am going back on what I just said in the introduction. I am not. What I think is that the best kind of strategy for establishing compatibilism is one which provides reasons to reject each incompatibilist argument that are specific to that argument. What I also think is that the only way for such a strategy to work is to argue from a well supported account of what it is to act freely and to be morally responsible for one's actions. Such an account would serve the function of unifying the specific responses to incompatibilist arguments and provide the resources to respond to any future incompatibilist arguments.

[17] I am not treating this as an oversight on their part, but rather as something they do not do because if they are correct one doesn't need to address the particular features of incompatibilist arguments.

this is the case, they do give evidence for an account of moral responsibility that is neutral between such positions.

Of course a full account of moral responsibility would tell us under what circumstances people are morally responsible for their actions. An account that complete could not be neutral between compatibilism and incompatibilism. To deal with this problem let me distinguish between two kinds of accounts of moral responsibility. A complete account of moral responsibility tells us, when conjoined with some factual information, when and why people are morally responsible for their behavior. It provides all the answers. Those involved in the central debates about moral responsibility are all working towards having a complete account of moral responsibility. As I said a complete account cannot be neutral with regards to the compatibility question, because it will answer the compatibility question.

A general account of moral responsibility is different. It is something that all contestants in the debates about moral responsibility can appeal to. It is what we are all talking about. Into a general account of moral responsibility can be stuffed all the truisms about moral responsibility, such as 'stones cannot be morally responsible for anything.' A general account of moral responsibility needs to be faithful to folk intuitions. What philosophers are talking about is the same thing that jurists and priests and other everyday people are talking about when they talk about moral responsibility. Most importantly a general account of moral responsibility is what is being appealed to by philosophers when they make arguments of the following kind 'Philosopher X's account of moral responsibility cannot be correct, because it violates a fundamental principle Y which is part of the concept of moral responsibility.' These kinds of arguments are common, but it is hard to see how could be convincing. After all Philosopher X has just said what he thinks moral responsibility is, and if that flouts principle Y, then that just means X doesn't think that Y is an essential part of moral

responsibility. At this point the philosophers quite often retreat into talking about their personal concepts of moral responsibility, and we find that what looked to be a disagreement wasn't one. This problem only comes up if we assume that in appealing to the nature of moral responsibility philosophers had to be referring to some complete account of moral responsibility. Since philosopher X is proposing a complete account, or the outlines of such an account, to presuppose some different complete account would be to simply assume X's view was false.

What we need is some shared concept of moral responsibility that does not give us all the answers, but which can be sometimes appealed to in order to rule out certain answers. A general account would be just this. Besides truisms, what content would a general account have? I do not want to take a stand on what any general account could or could not include, but I will be looking at some general accounts which consist mostly of claims about what we are doing when we hold someone morally responsible. I will argue that a good account of what we are doing when we hold people morally responsible will be neutral between compatibilism and incompatibilism. In other words, I will argue that a general account of moral responsibility that can be shared by all contestants in the debate is possible, and I will be arguing against Strawsonians who think that an account of what we are doing when we hold people morally responsible favors compatibilists.

1. Responsibility, Holding Responsible, Blame and Praise

One interesting thing about these Strawsonian accounts of moral responsibility is that none of them are explicitly put forward as an account of moral responsibility. Most Strawsonians attempt to give an account of what it is to hold someone morally responsible, or what praising and blaming someone amount to.[18] I think that in giving such an account, Strawsonians are giving an account of moral responsibility, or are

[18] See Wallace (1994)

coming close to giving one. These accounts of what holding morally responsible amounts to all contain commitments to what can and cannot make it legitimate to hold someone morally responsible. It is because of this feature of the accounts that they ought to be taken to be giving accounts of what it is for a person to be morally responsible. The reason for this is simple. It is very plausible that if it is legitimate to hold X morally responsible for Y, then X is morally responsible for Y. Endorsing this conditional claim does not commit the Strawsonians to abandoning the claim many of them defend; that holding someone morally responsible is prior, in some sense, to that person being morally responsible. What this priority claim amounts to is not always clear. It might be that holding responsible is evidentially prior to being responsible, so that we base our judgments about the latter on our knowledge of the former.[19] This evidential claim is clearly not imperiled by the claim that if it is legitimate to hold someone morally responsible then they are morally responsible. Alternately, the priority claim might amount to the claim that the explanation of why people are morally responsible is because they are legitimately held to be so. This explanatory priority claim is significantly stronger than the evidential priority claim, in that it suggests that what makes it legitimate to hold someone responsible is not some set of facts which includes facts about whether they are morally responsible. But even the explanatory priority claim is not imperiled by the conditional claim above. In fact it implies that conditional claim.

[19] This may sound paradoxical. Surely when we are deciding whether to hold someone morally responsible we consult our judgments about whether they are morally responsible, not the other way around. But it is not in the context of everyday decision making that this evidentiary claim is meant to hold. It is in the context of philosophical discussion that we take cases of holding someone morally responsible as evidentially prior to someone's being morally responsible. Because when and if people are morally responsible is precisely what is under debate we need to appeal to some more neutral body of data to help us decide between competing theories. On the assumption that people share the same concept of moral responsibility and are reasonably good at applying it, it makes sense to look at what actually prompts people to hold others morally responsible as a clue to when we think it is appropriate to do so. In other words this evidential priority claim is just the claim that our intuitions about when people are morally responsible are evidentially prior to any theoretical account of moral responsibility.

So, if it is legitimate to hold someone morally responsible then they are morally responsible. Is it true that if someone is responsible it is legitimate to hold them responsible? The explanatory priority claim implies that it is. The evidential priority claim does not. And anyone who thinks that an agent's status as morally responsible does not depend on how others may legitimately treat them or think of them will think that this second conditional claim is false. After all someone could be morally responsible for her actions while all evidence at a judge's disposal suggests they are not. In that case it is plausible that the judge may not legitimately hold the person morally responsible. So an account of the legitimacy of holding someone morally responsible will not, all by itself give us an account of what it is for a person to be morally responsible. But if you were to imagine an omniscient judge, whose evidence always supports the truth, you will be able to use the account of the legitimacy of holding responsible to get very quickly to an account of what it is for a person to actually be responsible for what she does. It is for this reason that I feel confident ascribing to Strawsonians various accounts of what makes a person morally responsible for what she does. It should be kept in mind however that some of the Strawsonians I will discuss might accept one of the two priority claims mentioned above.

In talking about what makes it legitimate to hold people responsible we learn a great deal about what it is for a person to be morally responsible. And to figure out what makes it legitimate to hold people morally responsible, we need to know what we are doing when we hold people morally responsible. This is a Strawsonian insight that I think is beyond reproach. So what are we talking about when we talk about our holding people morally responsible? We are talking about punishment and reward and

blame and praise, at least.[20] I think that of those activities, it is blaming and praising, and not punishing and rewarding, that are central to understanding moral responsibility. The reasons are fairly well known. Punishing and rewarding, because they aspects of general social management, are sensitive to facts that intuitively have nothing to do with the person being punished or rewarded or the action for which they are being punished and rewarded, and presumably moral responsibility has to do with the relation between people and their actions.[21] If you want to know why people are rewarded with millions of dollars for guessing the correct lottery numbers you will be told that the lottery system makes the state money that goes to fund education. Presumably the lottery winner is not being rewarded for her helping out schools, since the money she contributed by purchasing a lottery ticket is less than most other people contribute in property taxes, without getting the chance to be given millions. If you want to know why it is that there are certain sentencing guidelines for seemingly trivial narcotics offenses you might be told that without the threat of such sentences law enforcement would not be able to secure the cooperation of small time drug dealers or consumers in the investigations of drug cartels. The usefulness of a small

[20] When we blame and praise people we do not typically express that blame and praise, even to ourselves, by saying 'I hold you X morally responsible for that event.' But we sometimes do make these kinds of judgments. In formal settings, such as court proceedings, these judgments are explicitly debated. Do such attributions count as praise and blame? Not immediately because praise and blame include commitments to the specific moral quality of the action, and the bare attribution of moral responsibility need not do so. There is a worry about the bare attributions of moral responsibility that does not arise with regards to praise and blame, and it is precisely the formality of such attributions. Praising and blaming are unquestionably features of our ordinary thought and talk. To the extent that we think features of our ordinary thought and talk, or features of pre-philosophical or pre-theoretic thought, are of significant evidential value these bare attributions of moral responsibility seem suspect. If they tend to be made only in institutional settings we have to worry that the institution is set up as it is because of the influence of certain philosophical theories. This is unquestionably true of the justice system where philosophical theories, usually very old ones, have a significant grip on legal theorizing. For the purposes of this paper I will ignore these bare attributions of moral responsibility because of these suspicions.

[21] I am not denying that people can be morally responsible for things, like the consequences of their actions or the character traits which explain their actions, that are not actions of theirs. All I am claiming is that to be morally responsible for such things you must also be morally responsible for some action to which those consequences or character traits are related. I am also not claiming that the responsibility for consequences or character traits is somehow derivative, while moral responsibility for actions is more central.

time marijuana dealer in bringing down other, more hardened, criminals doesn't seem to have anything to do with the moral status of the drug dealer or the moral status of the action of selling marijuana.

A detailed account of when and why we punish and reward people is not likely to give us a clear picture of what is involved in holding people morally responsible. The legitimacy of praise and blame is not obviously sensitive to the same broad spectrum of public policy concerns that help determine whether and how much we punish or reward people.[22] That is why punishment and reward should not be looked to when we are trying to figure out what is going on when we hold people responsible.[23] So I am asserting an equivalence between holding someone morally responsible and blaming and praising them. More interestingly I am asserting:

> **Equivalence**: It is legitimate to hold X morally responsible for y-ing iff X is praiseworthy or blameworthy for y-ing.

It might seem that according to **Equivalence** a person cannot be morally responsible for actions that are neither right nor wrong. While I don't think that is true it is easy to see why someone might think it is. It is natural to think that a person cannot be praiseworthy for an action unless the action is right and cannot be

[22] This is not to say that actual habits of praise and blame are not sensitive to facts which are irrelevant to moral responsibility. It is probably true that people praise and blame women and racial minorities differently than they praise and blame men and racial majorities. But hopefully there is little risk of those differences being treated as legitimate, at least in the context of academic theorizing. The contrast with punishment and reward is that the legitimacy of punishment and reward is supposed to depend on issues that don't seem to have anything to do with moral responsibility, and with blame and praise this is not true.

[23] I want to mention one caveat, and that has to do with cases of horrendous crimes, such as war crimes or other crimes against humanity. The issue is that it sounds tremendously weak to talk about blaming genocidal dictators. It sounds as though we are debating about whether to slap on the wrist someone who butchered thousands or millions of innocent people. This is, I think, part of the explanation for why some people don't think we can actually hold the deeply evil morally responsible. It is also, I think, a mistake. The mistake lies in failing to recognize that punishment typically, though not universally as we have seen, involves holding someone morally responsible. Telling Hitler that he was a morally bad person is not the appropriate way to express to him that you hold him morally responsible, but that does not mean there is no way to express the fact that we hold him morally responsible.

blameworthy for an action unless it is wrong. While it is natural to accept both claims, neither is correct. Consider the following kind of case:

> *Inconsequential Corruption*: Tom, a functionary at the local county court, is offered a bribe by Rick, a well known street tough, to influence the judge of an upcoming case to let a friend of Rick's get off easy. Tom accepts the money and talks to the judge. In fact Tom has no influence over the judge and cannot possibly succeed at what he has been paid to do. Also, Rick is actually an undercover policeman who knows that Tom suffers from delusions of grandeur and cannot possibly get the criminal off the hook. Rick is just paying Tom to make it look like he tried to secure the criminal's release to other members of the criminal organization that Rick has infiltrated.

I think that in this case what Tom does is not wrong. He takes money from someone who is not aiming to do harm by giving him the money, but who is rather trying to do good. Tom is not going to get the criminal off the hook, nor is he going to do anything that is at all likely to result in getting the criminal off the hook. These facts are sufficient, I think, to show that unlike most cases of taking a bribe, Tom's taking the bribe is not wrong.[24] But the fact that Tom thinks this is a normal case of bribery, where he is going to get the criminal off the hook and the briber means for him to do just that, is sufficient for Tom being blameworthy for taking the money.

[24] It might be the case that Tom's deciding to take the bribe was wrong, so that there is something in *Inconsequential Corruption* that he is both blameworthy for and that is wrong. All I want to insist on is that there is one thing in the case that Tom is blameworthy for but is not wrong, and that is his taking of the bribe. Even if his taking of the bribe is blameworthy only because his deciding to take the bribe is blameworthy, and even if that decision is only blameworthy because it is wrong it is nonetheless true that X can be blameworthy for Y even if Y is not morally wrong, and so true that X can be morally responsible for Y even if Y is not morally wrong. *Inconsequential Corruption* is not a counter-example to the claim that X can only be morally responsible for Y in virtue of something's being right or wrong, but there is little reason to be troubled by this latter principle. Anyone who accepts the claim that actions have exactly the moral quality that the decisions which produced them have will of course think that Tom did something wrong, and so those who accept that claim will have to accept that **Equivalence** entails that we can only be morally responsible for actions that are right or wrong. Such views might be right but they incur the cost of seeming to not be able to make sense of someone behaving correctly for the wrong reasons. They can make sense of someone tokening an act type which is obligatory for the wrong reasons, but they cannot make sense of the act token being the right action and being done for the wrong reasons. So even though **Equivalence** when coupled with certain moral theories has the counterintuitive consequence I am trying to avoid, I think that **Equivalence** when coupled with intuitive (but of course corrigible) ethical claims is not.

A similar case can be presented to show that people can be praiseworthy for actions they perform that are not right actions. Consider:

> *Underinformed Voter*: Jane has moved to Pennsylvania in the fall. She moved early enough to be eligible to vote in the upcoming congressional elections. She wants to support the legal right to an abortion, and so wants to support the candidate most likely to support such rights. For the sake of argument suppose that supporting legal rights to abortion is the right thing to do. There is insufficient time in Jane's busy schedule to investigate the positions of the candidates in any depth. She knows that in general Democrats are more likely than Republicans to support legal rights to abortion, and so she votes for the Democrat. In this case however the Republican is in favor of legal abortion rights and the Democrat is not. So Jane ends up voting for a congressman who does not support the legal right to an abortion.

On the assumption that supporting the legal right to abortion is the right thing to do, Jane has failed to do the right thing. But because it was her intention to vote for someone who supported such rights, and because the decision she made was rational given the information she had, and because she did not have sufficient time to get the information that showed that her rational decision was incorrect, it seems to me that she deserves praise for what she did.

So **Equivalence** does not imply that we can only be praised for right actions or that we can only be blamed for wrong actions. Intentions matter for praiseworthiness and blameworthiness in a way that intentions do not always matter to the rightness or wrongness of actions. The suggested equivalence does have the odd consequence that actions that are neither right nor wrong and that the agent knows are neither right nor wrong, are actions for which the agent is not morally responsible. This is odd because it means that we are not morally responsible for tying our shoes in the morning, or for getting a cup of coffee. To alleviate the oddness of this suggestion let me point out that it is compatible with the suggested equivalence that we might still be causally responsible for tying our shoes or getting a cup of coffee. We would just not be morally responsible. It is also the case that we are not excused for tying our shoes or

getting coffee, despite the fact that saying 'X is not morally responsible for Y' is normally a way of excusing X for doing Y. In this case there is nothing to be excused from. Certainly there is an action, but it is of no moral significance, so there is no blame or praise to which one would be subject that one needs to get out of with an excuse.

2. Obstacle Clearing

2.1 Excuses and Exemptions

Strawson actually advances several different arguments, and not just the Emotion Account, which take claims about the nature of praise and blame as premises and yield conclusions about the tenability of incompatibilism. Because some of these arguments remain popular, I will show why none of them work.[25]

Strawson seeks to show that certain kinds of justification of our practice of holding people morally responsible are out of place. The kinds of justification that are out of place are sometimes labeled as 'rational', 'theoretical', and 'metaphysical'.[26] What was important to Strawson was to show that two views, one on which the propriety of our ascriptions of moral responsibility depends on their usefulness in controlling the behavior of others, the other on which it depends on the truth of some indeterministic theory of action, are confused. They are, according to Strawson, attempting to provide exactly the kind of justification that is out of place and uncalled for. The results of his approach are that one form of compatibilism, and all forms of incompatibilism are ruled out.

[25] Kevin Magill (1997) seems to think that these arguments work.

[26] Strawson says that both sides of the free will debate, and those he thinks are utilitarian compatibilists and libertarians, "seek, in different ways, to over-intellectualize the facts." They are guilt of this he says, because the practice of moral responsibility "neither calls for, nor permits, an external 'rational' justification." See Strawson (1974) pg. 23 In a similar vein he says that "the human commitment to participation in ordinary inter-personal relationships is, I think, too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical conviction might" remove it. (Strawson (1974)pg.11) Strawson derides the incompatibilist for his "panicky metaphysics" (Strawson (1974) pg. 25) and

Strawson presents two arguments which can be taken as straightforward arguments for compatibilism, and they both fail. To present the first I will need to introduce some terminology.[27] An *excuse* is a withholding of blame[28], where it would normally be warranted, on the basis of some feature of the action which marked it as somehow not indicative of ill-will. An *exemption* is a withholding of blame where it would normally be warranted, on the basis of some relatively stable, but abnormal, feature of the person involved, which made them an inappropriate target for the reactive attitudes and moral judgment.[29]

By way of example, if someone steps down hard on my foot that would normally be a reason to blame them. But if I learn that the person who stepped down hard on my foot did so because they had been pushed and lost balance, not only would there be no reason to blame them, blame would in fact be illegitimate. The reason is that they did not intend to hurt me, and so in hurting me did not express ill-will towards me.[30] In this case I excuse their stepping on my foot. Now if, after I excuse them for stepping on my foot, I find out that the person robbed a bank and killed two people in the process, I have no reason to not blame them. The facts which show that an excuse is warranted do not show that any future blaming is unwarranted.

That last feature is what makes excuses importantly different from exemptions. Imagine the same situation as before, where someone has stepped down hard on my foot. If, instead of finding out that the person was pushed and off balance, I find out that they are mentally disabled, and mentally disabled enough that they do not realize that stepping down hard on a person's foot hurts them, or do not realize that they have

[27] Originally introduced by Gary Watson (1987).
[28] I am only talking about blame here because it is not natural to express what it is to excuse someone in terms of praise. I think Watson and Strawson must have thought there was an analogous distinction to be made between cases of withholding praise.
[29] This terminological distinction is based on the categories Strawson sets up in section IV of Strawson (1974)
[30] This is Strawson's account of why blame is not appropriate in these cases.

stepped on my foot at all, then I have reason to withhold blame. But unlike the first case, this withholding extends to all the person's actions. If this mentally challenged adult goes on to kill two people, I will have reason to withhold blame, on the basis of the very same facts which led me to withhold moral judgment when he stepped on my foot. I exempted them from blame upon learning that they were mentally challenged, and so long as they retain this feature they will merit such an exemption.

2.2 The Abnormality Argument

In addition to the scope of the withholding being different, excuses and exemptions differ in the kinds of reasons which support them, according to Strawson. Excuses are merited because the action for which the person is excused does not actually express the ill-will that instances of that action type usually do. The reasons for exemptions are more complicated, as will be clear after the discussion of the following argument:

> **Abnormality of Exemptions[31]:**
> 1. Determinism does not imply anything about the content of our mental states.
> 2. So it does not imply that we do or do not express good will in all situations.
> 3. So determinism does not give reason to excuse every action.[32]
> 4. We exempt people only because they are abnormal.
> 5. If determinism implied that we ought to exempt everyone from moral judgment then determinism would imply that everyone was abnormal.
> 6. It is logically impossible for everyone to be abnormal.
> 7. So if determinism is logically possible then determinism does not imply that everyone is abnormal.
> 8. So if determinism is logically possible then determinism does not imply that we ought to exempt everyone from moral judgment
> 9. Excuses and exemptions exhaust the ways in which one can be bound to withhold moral judgment.
> 10. So if determinism is logically possible then determinism does not

[31] See Strawson (1974) pgs. 10-11

[32] Calling this entire argument the 'Abnormality Argument' is an admitted misnomer on my part, as premises 1-3 form a sub-argument that is both essential to the overall argument and does not mention abnormality. I call the entire argument 'Abnormality' because it is only the claims about abnormality that are of interest as far as I am concerned.

imply that we are always bound to withhold moral judgment.
12. So either determinism is logically impossible or determinism does not imply that we are always bound to withhold moral judgment.

I do not wish to commit myself to the argument for (3). There might be reasons to think that determinism does have something to say about the content of our mental states; specifically one might think that if determinism is true then no mental state could count as being morally worthy.[33] But perhaps by 'good will' Strawson merely means something like a will that wishes you well. If so then there is no good reason to reject (3). I do think that the argument for (8) fails. The problem is with premises (4) and (6). Specifically there is no sense of 'abnormal' on which both (4) and (6) are true. What one might normally mean by 'abnormal' is something like 'statistically unlikely' or 'rare'. On this meaning (6) is true. It is logically impossible for everyone to be rare, and logically impossible for it to be statistically unlikely that something is true which is such that if it is true it is true for all things and for all time. But on this meaning (4) is not true. Abnormality in this sense is clearly not sufficient for the withholding of moral judgment. We do not withhold judgment for the wise or the very virtuous, though they are rare. Abnormality in this sense might be necessary for the withholding of moral judgment. The cases that Strawson mentions as cases of exemptions include childhood, mental disorders like schizophrenia, systematic mental perversion and compulsive behavior.[34] These cases are all (excluding childhood), hopefully, abnormal cases in the sense under consideration. But for none of these cases is it true that we withhold judgment on the basis of abnormality.

To see this consider a possible world in which one of these characteristics is actually common. In that possible world something like internal compulsion is not

[33] Kant seemed to think something like this, insofar as he thought the spontaneity of practical reason was a necessary condition of having a good will.
[34] Strawson (1974) pg. 8

rare or statistically unlikely.[35] Perhaps a plague has swept through the populace and affected almost everyone's ability to resist temptation, such that they cannot rationally deliberate about their actions.[36] In this world let us suppose that those not affected by the plague withhold judgment of those affected by the plague. They do this because they realize that those affected by the plague lack rational control over their actions. This response is one we should approve of and recommend, and the upshot of this is that abnormality is not the reason why, when we withhold judgment on account of things like compulsion, we withhold judgment. When abnormality and certain mental conditions come apart, it is the mental conditions that withholding tracks, not abnormality.

There is a second sense of 'abnormal' that might be meant. 'Abnormal' could mean 'failing to reach or meet certain normative standards.'[37] In this sense it may be true that we withhold judgment on account of abnormality. Some have interpreted Strawson to claim that there are standards of normative competence that we must meet if we are responsible[38]. Whether this sense is sufficient for the truth of (4), it is clearly

[35] There is a potential wrinkle here. 'Abnormality' might be taken to rigidly designate. If it designates the set of behaviors which bear the property 'rare', then in this possible world, despite the fact that these behaviors are not rare, they are abnormal. Now I do not think it is plausible that 'abnormality' rigidly designates, but even if it does the result of this would be that it is not logically impossible for everything to be what is abnormal, because what is now abnormal might at some point not be rare or statistically unlikely. Not only is there some possible world where everyone is at some point abnormal, if it rigidly designates what is now rare, but that world might be the actual world.

[36] Paul Russell (1992) has presented a similar objection.

[37] I mean to deal here with claims that Strawson makes about lack of moral development. Strawson cites moral underdevelopment as a reason to withhold the reactive attitudes, and so he also thinks that determinism cannot imply that everyone is morally underdeveloped. This is just to beg the question against the incompatibilist though. The incompatibilist is going to want to say that a person's practical reason must work in a certain kind of way for them to count as a full fledged moral agent capable of deserving blame or praise, and that the way practical reason must work is non-deterministically. For Strawson to simply rule this out by fiat is a failure to actually engage the incompatibilist's argument. It would also count as begging the question if Strawson were to say that lack of moral development has to consist in a condition that prevents an individual from reaching some level of mental functioning that most other people meet. This is just to assume that the reasons for withholding the reactive attitudes includes a commitment to the claim that most people are morally responsible for what they do.

[38] Watson has said that being responsible for Strawson requires being able to take part in a moral community of reason-givers, and in his he has been followed by McKenna (2005). Wallace has also said that responsibility consists in the having of certain normative competencies.

not sufficient for the truth of (6). It is not clear why any normative standard would be such that not everyone could fail it.[39]

2.3 The Actual Judgment Argument

Strawson has a second straightforward argument for compatibilism, which I will call the Actual Judgment argument. The idea is that if we pay attention to what we actually take to be reasons, in actual cases, to withhold judgment we will see that determinism is irrelevant to moral judgment:

Actual Judgment

1. There is no sense of 'determined' such that:
   (a) If determinism is true, then all behavior is determined in this sense.
   (b) Determinism might be true.
   (c) Our withholding of moral judgment is the result of a prior embracing of a belief that the behavior of the human being in question is determined in this sense.[40]
2. If determinism is a threat to our ascriptions of moral responsibility, then it would fulfill the above three conditions.
3. So determinism is not a threat to our ascriptions of moral responsibility.

What is certainly true is that we do not actually take determinism into account in most deliberations about the casting of blame or deliverance of praise. But this is not enough to show that determinism is irrelevant to whether we should withhold moral judgment. It is possible that the reason that our withholding of moral judgment is never the result of a prior embracing of a belief that the behavior is determined is because we are confused about what agency requires. I tend to think that the nature of agency is not something that philosophers know about better than non-philosophers,

[39] This issue should be separated from the issue of demandingness in ethics. It may be that some system of duties are too demanding, and that this shows that they do not provide the correct moral standard. But not all normative standards are duties, and is not clear why we should be concerned that standards that are not duties should, under certain circumstances, fail to be met by everyone. There is a normative standard of perfection, and it is likely that no one meets it. It is not likely that the standards of normative competence are standards of perfection, but there is no reason to think that there cannot be any circumstances in which no one can meet them.

[40] These three propositions are found on Strawson (1974) pgs. 17-18

though perhaps we are more articulate on the subject. So I do not place much faith in the idea that I just mentioned, that people are generally confused about what agency is. I have no account to give of this knowledge, but I trust that most have it.

My trust is not enough to show that Strawson is right though. Even if we are not all confused about what agency requires, it is still possible that determinism is relevant. Since we do not actually take account of determinism in ordinary contexts, determinism is not immediately relevant to moral judgment. The thesis of determinism does not immediately contradict any claims to which we appeal when we deliberate about making a moral judgment, or even, we may grant Strawson, when we justify the decision to blame and praise, in normal contexts. It does not mean that the truth of the thesis of determinism does not rule out the truth of presuppositions we make when deliberating about whether to engage in moral judgment, or assumptions we generally accept when doing so.

It is plausible, initially at least, that we all accept a moral rule somewhat similar to 'Do not blame people for what they could not avoid.' Now the way we normally go about figuring out if someone could avoid what they did is by looking to see whether certain factors are at work. The most likely obstacles to someone being able to avoid what they do are things like interference from other agents, ignorance about what she was doing or whether she was doing it, or being under the influence of some kind of drug. Now it might be that all we do in normal cases is make sure such likely obstacles are not present, and then go on with blaming and praising confident that we have established that the person is morally responsible, or confident at least that we have done all we should be expected to in order to establish it. Determinism does not show that any of the most likely obstacles are present, and this is what Strawson is pointing out. But if we only care about the likely obstacles because of the aforementioned principle, 'Do not blame people for what they could not avoid', then

they cannot be all we care about. Any way in which someone could fail to be able to avoid doing what they in fact did would be something that we have to take account of. That we only normally look for certain obstacles to doing otherwise can be accounted for by the beliefs that they are the most likely obstacles and that it is only rational to look for obstacles at least as likely to hold as the set just mentioned.

Now either of these beliefs could be false. It might be that determinism is true, and so that every action is determined, and so, given some incompatibilist assumptions, that it is very likely that there is an obstacle to our avoiding what we do whenever we do anything. The argument I have just given is, in rough outline, a version of generalization strategy. Generalization strategies look to our normal reasons for withholding moral judgment, and try to infer from those reasons more general commitments.[41] With those general commitments, and the principle 'Do not blame people for what they cannot avoid' is one, in hand, the relevance of determinism can be evaluated with regard to those commitments. This strategy, often employed by incompatibilists, does not deny Strawson's claim that we do not actually withhold moral judgment on the grounds of determinism being true, or give it on the grounds that indeterminism is true. So incompatibilists have good reason, I take it, to accept (1) and reject (2).

3. The Emotion Account

3.1 Emotions, Cognitivism and Quasi-Realism

On to the Emotion Account. The general picture it suggests is this: the practice of holding people morally responsible is constituted by the reactive attitudes or our disposition to have the reactive attitudes, and because of this constitutive relationship there cannot be any reason to reject or let go of the practice of praising

[41] More will be said about generalization arguments later.

and blaming people.[42] So, in particular, determinism cannot be a reason to abandon the practice of blaming and praising, which it would have to do if it were incompatible with moral responsibility.[43]

What are the reactive attitudes? Strawson does not provide a strict definition of the reactive attitudes. He does provide an admittedly incomplete list, consisting of "gratitude, resentment, forgiveness, love and hurt feelings."[44] He identifies the reactive attitudes as the attitudes characteristic of "involvement or participation in a human relationship."[45] In contrast with the reactive attitudes Strawson presents the 'objective attitude', which is, if you adopt it towards another human being, to "see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be…managed or handled or cured or trained."[46]

These characterizations do not amount to a definition, and it is exactly a definition which some have called for.[47] Strawson says at one point that "It does not seem to me to matter if a strict definition is not to be had…given this characterization,

[42] This again is a very rough characterization of what various authors have said of Strawson, but again I think this principle captures the basic idea. See Watson (1987) "To regard people as responsible agents is to be ready to treat them in certain ways." (Pg. 256) and "In Strawson's view, there is no such independent notion of responsibility that explains the propriety of the reactive attitudes. The explanatory priority is the other way around: It is not that we hold people responsible because they are responsible; rather, the idea (our idea) that we are responsible is to be understood by the practice, which itself is not a matter of holding some propositions to be true, but of expressing our concerns and demands about our treatment of one another. These stances and responses are expressions of certain rudimentary needs and aversions…" (pg. 258) See also Wallace (1994) who says, "if we wish to make sense of the idea that there are facts about what it is to be a responsible agent, it is best not to picture such facts as conceptually prior to and independent of our practice of holding people responsible," (pg. 2) and "to hold someone responsible, I argue, is essentially to be subject to emotions of this class (resentment, indignation and guilt) in one's dealings with the person." (pg. 3) Bennett (1980) speaks of "the non-propositional nature of blaming and praising" (pg. 23-24)

[43] If determinism is incompatible with moral responsibility it certainly follows that the knowledge that determinism obtains gives us reason to abandon the belief that people are sometimes morally responsible. On the face of it this fact about what we have reason to believe implies that we have reason to stop blaming and praising people.

[44] Strawson (1974)) pg. 4

[45] Strawson (1974)) pg. 9

[46] Strawson (1974) pg. 9

[47] Wallace (1994) in particular, while Bennett (1980) treats it as a failing of the Strawsonian view, which is his view as well, that it cannot provide such a definition.

we can recognize that the responses which fall under it, or very similar responses, may also be evoked by behaviour which does not, or does not strictly, fall within its scope."[48] I agree and so will not try to give a strict definition. There will be points where knowing more about the reactive attitudes will matter, and when these issues are raised, I will deal with them. But it does not seem to me that the fact that Strawson cannot articulate what makes feelings like gratitude, love and resentment similar implies that he, or anyone else similarly unable, should be barred for talking about this class in a philosophy paper.

In addition to the initial list of attitudes, Strawson says there are "sympathetic, vicarious, impersonal, disinterested, or generalized analogues," of the reactive attitudes.[49] Unlike the reactive attitudes, which are responses to the presence or absence of goodwill other people have towards us, the analogs of the reactive attitudes are responses to the presence or absence of goodwill directed towards third parties. This includes both the way that other people treat third parties and the way that we ourselves treat third parties. As will be important later, Strawson, in talking about this set of reactive attitudes, is attempting to give an account of morality in general.

There are two things to get clear on. First, we need to know what exactly is being constituted by the disposition to have the reactive attitudes. To answer with 'the practice of moral responsibility' is rather vague, and I will offer two more precise formulations. The first formulation amounts to the claim that each instance of blame and praise is constituted by some reactive attitude.[50] The second formulation makes no claims about the status of instances of praise and blame, but says rather that there is a practice of holding people morally responsible, and that the rules governing that

[48] Strawson (1980) pg. 266
[49] Strawson (1974) pg. 14
[50] Bennett (1980) and Wallace (1994) support the Constitution thesis on this formulation. Whether or not Strawson meant the view to be taken this way I will stipulate that praising and blaming just are the occurrence of the relevant emotions, and that to express that one blames or praises someone is just to express those emotions.

practice is determined by our actual disposition to have the reactive attitudes. On this picture there is a practice, within which beliefs are formed and rejected and justifications accepted and debated, which at base is constituted by certain social habits of ours. These social habits do not provide reasons for the practice, but do figure in explanations for why the nature of the practice is as it is. While these social habits, the habits associated with reacting to people in certain emotionally toned ways, do not provide reasons, nothing else does either. Demanding a reason for a practice is, on this picture, a category mistake, as reasons can only show up within a practice.[51]

The second thing to get clear on is the nature of the reactive attitudes. It is clear that they are a species of emotion, so getting clear about their nature will involve getting clear, or clearer anyway, about the nature of the emotions.[52] Theories of emotions can be arranged on a spectrum. At one end of the spectrum is the view that emotions are cognitive[53]. While they differ from beliefs in being presented under a felt mode of presentation, they are like beliefs in almost every other way. They have not only intentional objects, but also propositional content, and truth-values. At the other end of the spectrum is the view that emotions are completely non-cognitive[54]. They are not beliefs presented under a felt mode of presentation, they are just feelings. They have no intentional objects[55], and certainly do not have propositional structure or truth values. Most theories fit somewhere in between, and there are few, if any,

[51] The two different ways of looking at Strawson's constitution claim are significantly different. On the first the constitution claim is a semantic claim, and on the second the constitution claim is not a semantic claim, but rather is a claim about what reasons there are and what reasons there could be that speak to holding others morally responsible.

[52] I am assuming that what distinguishes the reactive attitudes from other emotions is just the fact that they are directed towards people because of their actions. If this is right then getting clear on what emotions are or can be will amount to getting clear on what the reactive attitudes are.

[53] Nussbaum (2001), Neu (2000) and Solomon (2004) are examples of philosophers who take a strongly cognitive view of emotions

[54] It is not clear that anyone in the debate about the nature of the emotions takes this view. Others have, and it was probably the dominant view until the last 30 years.

[55] Those who subscribe to this extreme form of non-cognitivism would say that while emotions like anger are caused by certain circumstances, they are not directed at any set of circumstances. The only sense to be given to the claim that 'I am angry at him' is that he caused you to be angry.

contemporary theories occupying the fully non-cognitive extreme of the spectrum.

This way of getting clear about things seems to have yielded a glut of interpretations of what Strawson's position is. There are two ways to understand what is being constituted, and as many ways to understand what is doing the constituting as there are interestingly different positions on the spectrum of theories of emotions. Because of the multiplicity of ways of understanding the Constitution Thesis, and because no Strawsonian has said very clearly where they stand with regard to these questions, my arguments will have at best tentative conclusions. I cannot deal even with all current theories of emotions, and of course cannot deal with the new theories which are bound to pop up in a lively area of research. I hope, however, to cover the major theories, and hope that doing this will be sufficient grounding to make the following claims:

> 1. The view on which each instance of praise and blame is just the expression of a reactive attitude is either, depending on the account of the emotions that is correct, neutral between compatibilism and incompatibilism or simply implausible.
> 2. The view on which the practice of blaming and praising does not stand in need of any justifying reasons is one which, even if it can be made coherent and plausible, need not worry the incompatibilist, for their criticisms are best conceived of as internal to the practice, not as external criticisms which seek to undermine the justification of the practice as a whole.

I will start by looking at different theories of the emotions and what they imply. What would follow from the claim that instances of blaming and praising are just expressions of emotions, also assuming that emotions are basically just like beliefs? It does not seem to imply much of anything, or at any rate it does not imply anything that would be controversial to anyone in the debate about moral responsibility. After all, incompatibilists and compatibilists seem to be arguing about whether certain propositions are true or false. On the widely accepted view that beliefs are attitudes towards propositions, and that beliefs share the truth values of the

propositions towards which they are directed, incompatibilist and compatibilists are just arguing about whether the beliefs that in part constitute the reactive attitudes are true. Strawson's claim about the relationship between blaming and praising and the reactive attitudes, when conjoined with a cognitivist account of the emotions, seems as though it could only have the interesting implications Strawson takes his view to have if the beliefs that in part constitute the reactive attitudes can only be directed at propositions whose content is insensitive to facts about determinism.

That the beliefs or judgments that constitute emotions have distinctive content is actually the dominant view among cognitivists about emotion. That content is almost always thought to be evaluative[56]. Is there any reason to think that if moral judgments were expressions of beliefs that have evaluative content, that this would show that incompatibilism is false? The only reason that this could be true is if the propositions under debate between incompatibilists and compatibilist are all both not evaluative propositions themselves and do not imply any evaluative propositions. This is not plausible. Take, for example, the Principle of Alternative Possibilities (PAP). According to PAP, if X is morally responsible for y-ing, then X had some real alternative to y-ing. Now if PAP is true then what would be true of a situation in which X was punished for y-ing, when X had not alternative to y-ing? This would amount to punishing X for y-ing when X was not morally responsible for y-ing. What is clear is the justificatory role played by moral responsibility. Typically X's moral responsibility would be a necessary condition of legitimately punishing X. Perhaps sometimes consequentialist considerations can, all by themselves, justify punishment. Even in such cases, however, it is still going to be the case that X was treated unfairly. After all X was punished for something that X was not morally responsible for, and that is unfair. So, no matter whether the punishment is all things considered justified,

[56] Nussbaum (2001), Solomon (2004), de Sousa (1987).

there is a pro tanto reason not to punish X. And, it seems to me, if there is a pro tanto reason not to punish X and we do, then there is something regrettable about the fact that X was punished. So propositions with evaluative content are implied by standard incompatibilist principles.

I move now to the non-cognitivist extreme. Currently there are few, if any, philosophers working on the emotions who take the extreme non-cognitivist view of emotions, and so this might seem like a useless enterprise. It is not. While it is not common today to take this view of the emotions, it was at Strawson's time, and at least one major interpreter has explicitly endorsed the non-cognitivist view that the reactive attitudes are just feelings.[57]

A radical non-cognitivist about the emotions will want to deny all the key claims of cognitivism about the emotions. So they will certainly want to deny that an emotion just is a belief. They will also want to deny that an emotion is an attitude directed at a proposition in such a way that if the proposition is false, then the emotion is inappropriate. They might also want to deny that emotions have intentional objects at all. What a given non-cognitivist will deny will likely depend on what that non-cognitivist takes to be the central commitment of cognitivism. I think that at the very least, any non-cognitivist account of the emotions will deny that any emotion has propositional content. If this were the case then if the proposition were false, the natural conclusion to draw would be that the emotion is unjustified. The whole point, I take it, of the Strawsonian move towards thinking of blame and praise as expressions of emotions is to make it impossible for the truth of determinism to imply that blame and praise are inappropriate. Making the acceptability of the emotions insensitive to the truth of propositions generally would do that job, and the best way to do that is to

[57] Bennett (1980) seems to have adopted this view of the reactive attitudes, saying that "it seems clear that in his view reactive attitudes are to a large extent matters of feeling." (pg. 32)

make it the case that they are not attitudes directed at a proposition, in the way that a belief is.[58]

I think such a non-cognitivist view about the emotions is extremely implausible. One clue that it is implausible is that almost no one accepts it. In fact, despite the view that is sometimes attributed to him, Strawson almost certainly could not have accepted this view. To see why remember that Strawson endorses conditions of acceptability on the reactive attitudes. These conditions have to do with whether the action being reacted to expresses good or ill will on the part of the agent, or whether the agent being reacted to was capable of having a will that was good. So, if it is the case that X in y-ing does not express ill will towards some other agent, Z, then it would be inappropriate for Z to resent X, for example. But this just means that resentment depends for its appropriateness on the truth of propositions about the content of a person's will and the relations that content bears to her actions.[59] So I don't think that Strawson did or could have endorsed the radical non-cognitivist view about the emotions. And that is a good thing for Strawson, because this radical non-cognitivism is extremely implausible. To see why consider the following exchange:

> Husband: I am angry at you for cheating on me!
> Wife: But I didn't cheat on you!

[58] Emotions would still depend for their acceptability and appropriateness on the truth of some propositions. Consider the proposition that some emotions are acceptable and appropriate, or that none are. Outside of these propositions which are about the appropriateness of emotions it is hard to see what propositions could entail that emotions are or are not appropriate or acceptable.

[59] It might be that the emotions are appropriate iff certain propositions are true, but that these propositions are not part of the content of the emotions. After all some actions are appropriate iff certain propositions are true, but those propositions are not part of the action. It is only appropriate to skip the meeting if I have been told that I don't need to go. Skipping the meeting doesn't have content at all, much less propositional content. Why not treat emotions the same way? This would be for inappropriate emotions to be unjustified in the same way actions can be. But this doesn't capture the phenomenon correctly. It is not just that emotions are unjustified on occasion, which of course they sometimes are, but that they are in an important sense incorrect. It is not just wrong for the husband to be angry, he is making a mistake. Either he is ignorant or he is being irrational (in some less robust sense of 'rational' than the Kantians have in mind), not just being unfair or inconsiderate. This is not a decisive consideration of course, but I think the burden of proof should be on the person who thinks that the truth of propositions determine the appropriateness of emotions without propositions being part of the content of those emotions.

Husband: I never thought you did!
Wife: Then you shouldn't be mad at me!
Husband: Don't tell me how to feel, I can feel anyway I like!

The problem here is that it is obvious that the husband's anger is inappropriate, irrational and unfair if what he is angry about, his wife's infidelity, never happened. But it is hard to see how, on the radical non-cognitivist account of the emotions, facts like 'she never committed adultery' could entail that his anger is inappropriate, irrational or unfair.

The challenge for the Strawsonian who thinks that blame and praise are expressions of emotion and who also thinks this fact shows that determinism cannot be incompatible with moral responsibility, is to find an account of the emotions which both allows what Strawson allows, that facts about the good or ill will of others can make the reactive attitudes appropriate or inappropriate, but disallow what the incompatibilist asserts, that determinism's being true makes the reactive attitudes inappropriate. We have already that seen that a simple cognitivist view, a cognitivist view on which the propositional content of emotions is limited to evaluative content, and a radical non-cognitivist view won't meet this challenge. Are there other views of the emotions that can do better?

Such a view would have to admit that emotions have intentional objects, and perhaps propositional content, but also show that those objects and that content are not connected to the issue of determinism in any way that would let incompatibilist concerns get a foothold. One might think that the somatic theory of the emotions, according to which emotions provide information about our own bodily states, is a good place to look[60] On the classical expression of this view, it is awareness of facts about our body like the fact that our heart is racing, that we are trembling, that we are crying, that our breathing has quickened, etc. that prompts emotional responses. How

[60] First suggested by William James (1884). Defended recently by Damasio (1999) and Prinz (2004).

is this not just a version of the radical non-cognitivist view? Well, for one thing, on the better versions of the somatic theory emotions are essentially perceptions of bodily change. Perceptions stand in relations to propositions that non-cognitivists about emotion deny emotions stand in. For example, if you have a perception of a stick being bent in water, and it is false that the stick is bent, then your perception is not veridical. Further, any judgments you base on that perception, the judgment that the stick would break with but a little pressure for example, would be unjustified. So on the somatic theory emotions can be veridical or non-veridical, just as other perceptions can be. And emotions can provide evidence for other judgments, just as perceptions do.

On this description of the somatic theory of the emotions the proposition that determinism obtains does not seem like it could be such as to entail that any emotion was veridical or nonveridical. After all that proposition is not a proposition about states of the a person's body. The problem for Strawsonians is that facts about X's will are also not facts that Z's emotions could be sensitive to. It might be objected here that facts about another person's will, while not themselves about our physiology, can easily cause the kinds of physiological change to which the emotions can be sensitive. This is certainly true. That someone is insulting me makes the color of my face change, as well as raising my heart rate and the temperature of the skin in my cheeks. If we were to alter the somatic theory ever so slightly so that emotions are parts of perceptions, but that the bodily changes to which they are reactions are also parts of those perceptions, then it would be the case that in being resentful or grateful we might be responding the ill or good will of others. Does this slightly altered somatic view help the Strawsonians?

No, because there is no reason to think that facts about the psychology of other people are any more likely to cause physiological changes in us than more theoretical,

metaphysical facts might. For example, if you are a Wall Street trader and you look up from your papers to see that the New York Stock Market has plunged 600 points in an hour, this fact will produce immediate physiological changes. You will perspire, you will get dizzy, you will experience, in short, all the physiological changes standardly associated with fear. But it will only have this effect because of some pretty advanced theoretical knowledge you have. Similarly, if you have been a theist your whole life, and also been committed to theism precisely because you accepted something like the Design Argument for God, becoming convinced of the Theory of Evolution is going to produce some pretty strong physiological changes in you. There is no reason, in principle, to think that knowing that determinism is true could not produce similar reactions, and so no reason to think that determinism cannot be relevant to whether or not certain emotions are veridical or not.[61] Abstract theoretical facts can have important consequences for matters of everyday life which can prompt certain physiological responses. If, as some have suggested, determinism is incompatible with our having a special metaphysical status as agents, then learning that determinism is true could prompt us to feel sorrowful, in the same way that learning that the love of our life no longer enjoys our company could. The somatic theory, in short, does not give any reason to think that the truth of determinism cannot make having the reactive attitudes irrational or incorrect. All that could do that job would be to show that determinism is not relevant to anything we care about, because if it is relevant to something we care about, then knowing that determinism is true would and should prompt certain emotional reactions. But to successfully argue that determinism is not relevant to anything we care about one would have to establish that compatibilism about determinism and moral responsibility is correct, along with some

[61] Of course if determinism was true, but one didn't know that it could not produce the emotional reaction. But then the absence of the emotional reaction would just constitute ignorance. So one would either be irrational or ignorant in continuing to have the reactive attitudes in such a case. This is not a conclusion unfriendly to incompatibilists.

other controversial doctrines. So appealing to a somatic theory of the emotions cannot establish that compatibilism is true.

This does not exhaust the potential views of the emotions though. I earlier ruled out radical non-cognitivism about the emotions as implausible and unusable for Strawsonians. What about a less radical non-cognitivism though? For example many philosophers have thought that simple emotivism about moral expressions is implausible. But emotivism is not the only form of non-cognitivism about moral expressions, and radical non-cognitivism is not the only kind of non-cognitivism about the emotions. Strawsonians might make appeal to varieties of non-cognitivism that are more plausible than the simple variety I have discussed. I am speaking of the general project of making emotivism about moral expressions compatible with as many features of ordinary moral discourse as possible, and I will take as example of this project the quasi-realist project of Simon Blackburn's. Essentially quasi-realism starts from a rather rigid distinction between the propriety of the intellectual or linguistic operations characteristic of a discourse, and the theories about those intellectual or linguistic operations. Anti-realism, Blackburn says, is a theory about the nature of those operations, and, in virtue of the distinction just mentioned, it tells us nothing at all about the propriety of the operations. The quasi-realist is just the anti-realist whose project it is to maintain anti-realism while figuring out a way to imitate or mimic the intellectual or linguistic operations of a discourse that the realist endorses. If this project is completely successful, such that the quasi-realist successfully finds a way to mimic the realist, then, Blackburn says, there is no important difference between realism and anti-realism. They are two theories that have the same upshot or cash value, differing only in the size of the ontology that they demand.

There are three conditions for the success of the quasi-realists project with

regards to discourse X. The first is that the following proposition is true; 'The only way for realism about area of discourse X, to be theoretically superior to anti-realism would be if realism implied or was consistent with the propriety of some intellectual or linguistic operation characteristic of X that anti-realism is inconsistent with.' The second is that the following proposition, expressing the rigid division, is true; 'It is not analytic that discourse X presupposes any realist or anti-realist ontology, or realist or anti-realist semantics.' The third is that the quasi-realist actually be able to say and think all the things that the realist can with regards to discourse X.

It is not my purpose here to argue against Blackburn, anymore than it is my task to argue for or against a non-cognitive account of the emotions. What I am interested in is what the implications of Strawson's view with regard to the tenability of incompatibilism would be, if we imputed some form of quasi-realism to him. What quasi-realism allows is that individual ethical judgments can be true and false, that they can stand in implication relationships with other ethical judgments, and so that they can justify actions and beliefs. If successful, quasi-realism does this while showing that there is no need for justification of the discourse by claiming that it represents the world or tracks real objective properties in it. The discourse itself is at base constituted by the having of certain attitudes and certain relations among those attitudes.

What is interesting is that if you impute a quasi realist account of blaming and praising to Strawson the view would look a great deal like the view he merely hinted at in "Freedom and Resentment", and endorsed explicitly in "Skepticism and Naturalism". The significant differences are that Strawson is talking about standpoints and practices and not discourses, and that Strawson is not primarily interested in

offering a semantic theory.[62] It is to this view that I now want to turn. What gives the view a chance of achieving Strawson's purpose is that he posits a relativism of practices, or standpoints that a person can occupy.[63] The first standpoint Strawson consistently describes as the participant or interpersonal standpoint. It is the standpoint we occupy because we are social beings, and when we are engaged in genuine interaction with other people. It is only when we are looking at the world from this standpoint that we are prone to the reactive attitudes. The second standpoint he describes as the 'objective' viewpoint. This is not supposed to signify that it is the second standpoint rather than the first from which we are able to see things as they are, in fact he denies this. 'Objective' here refers to the way we look at other people, as natural objects to be investigated and manipulated.[64] The objective attitude is supposed to the attitude we take when we consider the world as scientists do.[65]

[62] Though Strawson's view in Skepticism and Naturalism would if it were fully worked out, require some significant semantic commitments to make sense of how claims from the scientific and personal points of view don't conflict even when on the surface they appear to.

[63] Strawson (1985)says "I have spoken of two different standpoints from which human behavior may be viewed: for short, the 'participant' versus the 'objective', the 'involved' versus the 'detached'. One standpoint is associated with a certain range of attitudes and reactions, the other with a different range of attitudes and reactions. [These] Standpoints and attitudes are not only different, they are profoundly opposed. One cannot be wholeheartedly committed to both at once…How natural is it, then, to ask the question: 'Which is the correct standpoint? Which is the standpoint from which we see things as they really are?'…What I want now to suggest is that error lies not [on] one side or the other of these two contrasting positions, but in the attempt to force the choice between them…I say that the appearance of contradiction arises only if we assume the existence of some metaphysically absolute standpoint from which we can judge between the two standpoints I have been contrasting. But there is no such superior standpoint – or none that we know of; it is the idea of such a standpoint that is the illusion…We can recognize, in our conception of the real, a reasonable relativity to standpoints that we do know and can occupy." (pg. 38)

[64] Bennett says that the objective attitude is marked by teleological investigation (pgs. 36-40) Strawson (1974) says of the utilitarian account of moral judgment, which is just that of judgments as manipulations guided by scientific investigations of human nature, that it is "is painted in a style appropriate to a situation envisaged as wholly dominated by objectivity of attitude." Pg. 21

[65] Strawson (1985) from time to time refers to this standpoint as that of 'scientific realism' or 'reductive naturalism', and seems to think that the objective viewpoint is the viewpoint of a person who takes everything there is to be reducible to something physical or something describable in the vocabulary of physics. What has become clear since Strawson wrote, and to some extent before, is that one can be a naturalist without being a reductive naturalist, and one can certainly be a scientific realist without thinking that all scientific disciplines reduce to physics. It is not clear to me to what extent these false equivalences actually affect Strawson's argument. It may be that the reason he thought that from the scientific realist standpoint moral notions so clearly did not apply was that we limited himself to only the kinds of descriptions of actions and events which are licensed by physics.

Importantly the objective attitude is the only attitude which is warranted when we view the world that way, according to Strawson.

How does the relativizing move work? It is hard to say. I cannot make sense of the notion that a tendency to feel certain things could constitute a way of seeing the world. If anything our feelings would be caused by the way we see the situation, and would be dependent on that way of seeing. If emotions are beliefs then it would make more sense to say that they stand as the foundation of a way of seeing the world, especially if they are a kind of perceptual belief.[66] If we interpret the reactive attitudes as some kind of belief, perceptual or otherwise, then Strawson's standpoint relativism says that what it is to be in those standpoints is to accept certain beliefs. That he then goes on to say that these beliefs, despite being mutually exclusive, do not contradict each other raises a problem. It is hard to imagine how beliefs, one of which says 'X can bear moral properties' and the other which says 'X does not bear moral properties' do not conflict, and it seems that Strawson is saying something close to that. Just saying that there is no metaphysically favored standpoint from which to evaluate these beliefs seems, at best, to show that we cannot know which is true, not that they do not conflict.

I wish to put aside this issue and proceed on, because while I do not understand how proneness to certain feelings could constitute a standpoint, I do not have an argument that it could not.[67] Supposing Strawson is entitled to claim that the standpoint and practice of moral responsibility is constituted by our proneness to the reactive attitudes, it remains to be seen whether metaphysical considerations are irrelevant to moral responsibility. The way that Strawson's relativizing move works,

[66] de Sousa (1987), Rorty (1980)

[67] It is likely that the answer is that emotions would have to be something short of beliefs, and something more than feelings. For more on this see Zimmerman (2001). Zimmerman is concerned with how particular attitudes could displace with disqualifying, but the conception of emotion he employs seems as though it might go a long way towards making sense of the role of the reactive attitudes in constituting or grounding standpoints.

I am claiming, is that it allows that particular attributions of moral responsibility are the kinds of things which can justify and be justified. They can be justified because they have propositional content, and they can be justified in the same way that other claims can be, by looking to see whether the world is the way they say it is. Strawson, as discussed earlier, laid out the reasons one could have to excuse or exempt someone from moral judgment, and so correct attributions of moral responsibility require that the case in question is one where neither excuses nor exemptions are appropriate. They can justify for the same reason that attributions of moral responsibility justify actions in any theory of moral responsibility, because if someone is morally responsible it is appropriate to blame, praise, reward and punish. That is just what the concept is.

However, if what I said before is right, then the normal excuses and exemptions can be taken to generate broader reasons than those typically given to exempt and excuse people. In particular if incompatibilists are right then our normal practices of moral judgment generate reasons to excuse or exempt everyone if determinism is true. The reasons that incompatibilists give are best thought of not as external criticisms of the practice or standpoint of morality, but as reasons which are internal to that practice. They are after all moral reasons.[68] What reason can Strawson give to resist such an argument?[69] I take that it is at this point that Strawson's practical arguments comes into play. Generalization strategies should be rejected because there are practical limits to how much we could or should alter our normal

[68] It is pretty clear that Strawson only means there to be two standpoints, one of which contains all of morality. Someone might want to claim that moral concerns are external to the practice of blaming and praising, but I do not see how one could make that plausible.

[69] Kevin MaGill (1997) has claimed that any internal challenge to Strawson's arguments fail, because any other moral principle offered against principles like 'guilty people deserved to be punished' would necessarily have be a restricted moral principle, so that it in fact could not conflict at all. As far as I can tell there is no argument for this claim in MaGill. It is also hard to imagine what kind of argument there could be for this claim that would require going through the piecemeal process of defending compatibilism that I presented earlier, that of giving specific reasons to reject each uniquely incompatibilist principle.

practice.

3.2 The Idleness Argument

Strawson offers two pragmatic arguments, one of which he thinks is far more important than the other.[70] That argument, the idleness argument, essentially says that the role the reactive attitudes play in our social life makes it the case that we cannot stop adopting the reactive attitudes, and so cannot stop holding people morally responsible.[71] There are two ways to interpret this argument that I will suggest. The first point claims essentially that we are hardwired to take the reactive attitudes, and that any attempt to change our habits in this area will fail.[72] The second points to the role the reactive attitudes play in our conceptual scheme, and says that the collapse of the one brings with it the collapse of the other.[73]

Against the first position it is worth pointing out that there is a gap between something being natural and something being justified. It certainly does not follow from something's being such that we are naturally prone to it that it is justified[74], and

[70] Strawson (1985) says that 'What I was above all concerned to stress [in Freedom and Resentment] was that our proneness to reactive attitudes is a natural fact, woven into the fabric of our lives, given with the fact of human society as we know it, neither calling for nor permitting a general 'rational' justification." Pg. 265

[71] Strawson (Strawson (1974)) pg. 13

[72] In Strawson (1985)he treats this argument as descended from Hume's positions with regard to skepticism about the external world. I will call this version of the argument the Humean version.

[73] Strawson (1985) hints at this argument (pg. 13) and discusses Wittgenstein, who adopted a similar position with regard to skepticism about the external world. Strawson treats Wittgenstein's position as essentially the same as Hume's, though the difference is I think clear. I include this Wittgensteinian argument in the section about practical arguments, even though it is not clearly a practical argument at all, because Strawson treats it as a variant of his own idleness argument. I will not discuss it much because I do not see how it could be made to work. Wittgenstein's argument is that certain propositions are such that they make possible rational thought about anything, and so that there is no non-self-undermining argument against those propositions. But the claim that we are justified in holding people responsible for what they do is not such a proposition. We can still use logical arguments and engage in empirical inquiry even if we reject that claim.

[74] Kevin MaGill (1996) has attributed a different argument to Strawson than the one I have. According to MaGill what Strawson is assuming that to think that X stands in need of justification is to think that we could stop X if it turned out that it is not justifiable. So the naturalness of X is not meant to imply that X is justified, but rather that X need not be justified. I see little to recommend the assumption that to think that X stands in need of justification is to think that we could suspend X. The goal of the search for justification is in the first place to determine whether something is justified, not to change behavior. Typically that determination will be employed in changing the behavior, and will be taken as a reason to change the behavior, but this is not a necessary condition of deciding that X is unjustified.

it is not clear, without some extra assumptions, that something's being such that we are naturally prone to it gives any reason at all to think it is justified.[75] Given this, the argument sounds as though it is suggesting that because we cannot stop ourselves from adopting the reactive attitudes, it is not worth looking into whether we should stop it.[76]

A brief argument for this position would be that it is irrational to try to do something that we know to be impossible. If it is impossible to do anything about our commitment to the reactive attitudes then it is irrational to try to do anything about it. A further necessary assumption is that inquiries into the justification of a practice are, or are when they have a point at all, directed at either changing a practice or protecting it from being altered by skeptical considerations. Since there is no change to be had either way, inquiries into justification are pointless.

This argument has been accused of anti-intellectualism. A way to see why this is true is to see that this argument leaves a crucial effect of inquiries into justification out of the function of such inquiries. Finding out that something is unjustified might not be enough to alter our commitment to it, but it is certainly enough to change our beliefs about it. There is no good argument for the claim that our belief that holding people is responsible is a justified practice is itself unalterable. There are after all, some people who actually believe that we are unjustified in holding people morally responsible.[77] The only move left here, that I can see, is to claim that changes in such beliefs are unimportant, and this does seem to express an anti-intellectualism. If it is true that our practice is unjustified, presumably it is better to know that, or at least it

[75] Some have suggested that men are naturally prone to rape. No one seriously thinks that in offering this claim that evolutionary socio-biologists are, or take themselves to be, giving reasons why rape is justified, or reasons why rape does not fail some legitimate demand for justification.

[76] From Strawson (1974) "…it is useless to ask whether it would not be rational for us to do what it is not in our nature to (be able to) do." Pg. 18

[77] There is an important dissimilarity here between skepticism about the external world and skepticism about moral responsibility. With regards to skepticism about moral responsibility what is plausibly unavoidable is suspending the practice which is in question, while with regards to skepticism about all knowledge what is plausibly unavoidable is both changing our beliefs and the practices which are only rational given the truth of those beliefs.

could be. By way of example, consider the notion of human depravity. Some people believe that human beings are unalterably depraved, while most disagree. A Strawsonian position with regard to this controversy would be that if we are unalterably depraved, then we could do nothing about it, so there is no sense asking about whether we are so depraved. But presumably it is worth knowing that there is such a thing as human goodness, or that there is not, even if we cannot improve people if they are depraved. I cannot imagine taken such a blasé attitude towards this subject, and I cannot see the difference between that attitude with regards to depravity, and Strawson's suggestion about how to think about the problem of moral responsibility.

3.3 The Pragmatic Argument

While he never admits that the Idleness argument does not work, Strawson does present an alternate practical argument, which I will call the Pragmatic Argument.[78] According to the Pragmatic Argument, the only way to decide whether it would be rational to give up the reactive attitudes would be if it made people's lives better.[79] Given that the reactive attitudes are an essential part of interpersonal relationships,[80] it is hard to see how rejecting them could improve our lives, if they carry along with it all the joys of interacting with other people.

Strawson says that the rationality of the practice of holding people responsible can only be judged according to the gains and losses of human life. He then only mentions the fact that without interpersonal involvement, life would be much poorer. Now, while incompatibilists do not always present their views this way, the

[78] Strawson presents Carnap's response to skepticism in Strawson (1985), in such a way that the affinities between that view and his come through. Carnap's position is essentially that since there is no way to verify the existence or nature of the kinds of facts commonly taken to ground our practices, then any claims about them are meaningless, and so they cannot be said to exist. If we cannot cite facts to the effect that people are X, to justify our attributing X to them, then all that is left are practical reasons.
[79] Strawson (1974) pg. 13
[80] Something that Wallace (1994) questions. There is a sticky point here about the proper definition of the reactive attitudes, which I will not deal with because Wallace's criticism would simply be another way of showing that the Pragmatic argument doesn't work.

incompatibilist arguments can be seen as moral arguments. The argument is that it is not fair to hold people morally responsible, and thus open them up to punishment and blame, if they could not avoid what they did. That moral claim conjoined with the non-moral claim that determinism precludes anyone avoiding anything they do is enough to generate the incompatibilist claim that if the thesis of determinism is true then we should abandon our practice of making the kind of moral judgments involved in blaming and praising.[81] Because moral reasons are practical reasons, Strawson's argument, when the moral nature of incompatibilism is made clear, amounts to the claim that utility outweighs concerns of fairness.[82]

A second reason to reject the Pragmatic argument is that according to Strawson's Idleness argument, we cannot actually abandon the practice of moral judgment anyway. So the extreme costs that Strawson points to would not be incurred if we were to come to the judgment that the practice was not justifiable. If we came to that conclusion we could not actually abandon the practice anyway. What we could do is tinker with it, perhaps removing any punishments which could only be justified by appeal to notions of desert, and not rehabilitation or public safety. The rest of our lives we would be forced to maintain our commitment to an unjustifiable practice, and that would surely bring negative consequences along with it. No doubt we would feel guilty for blaming people, despite being unable to stop doing it completely. Strawson could claim that this cost was sufficient to give us good practical reason to not seriously entertain the question, and to proceed along with the assumption that our practice was justified. This is in essence the position that we should assume that the

[81] Strawson thinks that this subclass of moral judgments stands or falls with the practice of moral judgment generally. This is not essential to his view, and I don't know of a place where he offers an argument for it.

[82] I take it that this is a problem for any view. Many of the reasons people give for rejecting act utilitarianism is that it seems to have just this consequence (though of course many utilitarians would claim that they are simply giving a different account of fairness, rather than claiming that utility trumps fairness), for example.

answer to any potentially troubling question is the answer which would trouble us the least. There is nothing to be said for such a position.

4. The Relationship Account

4.1 Watson's Moral Community

It might be that I have been uncharitable to Strawson. While it is certainly true that he offered the Practical Arguments I have presented, and it is certainly true he offered the **Abnormality** and **Actual Judgment** arguments, it might be that he never meant to endorse what I call the Emotion Account. Gary Watson has presented an account of moral responsibility that I call the Relationship View which he takes to be the view that Strawson accepted, made more precise. I am prepared to leave it an open question whether Strawson accepted the Relationship View. What interests me more are the merits of the view itself as it has been presented by Watson and more recently by Tim Scanlon. These views, like the Emotion Account, look like they have direct compatibilist implications. I will argue, however that the Relationship Account is very likely false, and that the only way to revise it in such a way as to make it plausible is to revise it in such a way as to make it neutral between compatibilism and incompatibilism.

I am going to discuss Watson first, in part because his view is closer to Strawson's, as it is a development of Strawson's view in the face of some of the very problems I raised.[83] Watson seeks to give a unified account of exemptions. About exemptions Strawson says that they come in two kinds "of which the first is far less

[83] Watson says "It would seem that many of the exemption conditions involve explanations of why the individuals display qualities to which the reactive attitudes are otherwise sensitive. So on the face of it, the reactive attitudes are also affected by these explanations. Strawson's essay does not provide an account of how this works or what kinds of explanations exempt." (Watson pg. 228) This is very close to my criticism of Strawson that there is no account of the emotions that makes it the case that we can be expected to withhold praise and blame because of some explanations of action and that we cannot be expected to withhold praise and blame because determinism is true. Watson fears that without an explanation of what unifies the class of exemptions, it will be impossible to resist the incompatibilist offering her favored explanation, whether it is the lack of ultimate sourcehood or the lack of alternative possibilities.

important than the second. In connection with the first sub-group we may think of such statements as 'He wasn't himself', 'He has been under very great strain recently', 'He was acting under post-hypnotic suggestion', in connection with the second, we may think of 'He's only a child', 'He's a hopeless schizophrenic', 'His mind has been systematically perverted', 'That's purely compulsive behavior on his part.'"[84] What could connect such a variety of reasons to exempt someone from being held responsible for her actions?

Watson seeks to answer this question by reference to what he calls the conditions of moral address. This immediately ought to suggest two questions. The first is 'conditions of what for moral address?' Watson might mean 'conditions for the possibility of moral address' or perhaps 'conditions for the rationality of moral address' or more minimally 'conditions in which moral address actually tends to occur'. Conditions of the first kind seem to be exhausted by the same conditions which make it possible to use a language and which make it possible to refer to moral properties.[85] If by 'moral address' Watson literally means 'addressing moral claims to people' then the conditions of the possibility of moral address are just the conditions of intelligibly articulating moral claims.[86] It involves sufficient mastery of the use of moral terms and would be exhausted by the conditions for mastery of a natural language. It is not clear at all how those conditions could explain the categories of exemption that Strawson lists.

What is much more in the Strawsonian spirit is to take the conditions of moral address to be the conditions in which we actually tend to morally address people. The

[84] Strawson (1974) pg. 24
[85] On the assumption, of course, that in making moral judgments we are referring to moral properties. This assumes cognitivism, something that Watson has always been at pains not to do. Keeping open the possibility of some kind of expressivism the set of conditions of the first kind are just the same as those required to express one's emotions, commands, or whatever else one thinks is being expressed in the making of moral judgments.
[86] If he doesn't mean this then I am not sure what he could mean by the 'possibility of moral address'.

problem with this approach is that it is not clear how we could get an explanation of what unifies the set of reasons that we give to exempt people that Strawson has offered. One way of looking at this is that, as holding someone responsible is a form of moral address, when done publicly, Strawson has already provided us with his account of the conditions under which we actually do address people morally, and what we need is some explanation of why it is that we do that. The project is to search for what justifies our exempting people as we do.[87] So it seems that the only conditions which matter are the conditions under which it is rational or reasonable to morally address people.

The justifying account that Watson offers is that the conditions under which it makes sense to address someone morally are not met in the cases of exemptions, though they fail to be met in different ways. Watson mentions two conditions; "One condition is that…the other must possess sufficient moral understanding; another is that the conduct in question be seen as reflecting the moral self."[88] The condition of moral understanding is meant to explain why we do not hold children or the schizophrenic responsible for their actions, while the condition of reflection of self explains why we do not hold people responsible for uncharacteristic actions or actions performed under hypnosis.

Why do I call Watson's view a Relationship Account? The reason stems from the condition of moral understanding. "It is tempting," Watson says, "to think that

[87] It is possible of course that what justifies such a pattern of exemptions, if anything, will not be the same thing which actually causes us to so exempt people. Perhaps we are led to exempt people because of the internalization of social standards of behavior which we only internalized because of fear of social sanction, and which only have the content they do because of accidental historical facts. Now if we do find a justifying account of exemptions then there is no reason to think that such a justification could not become part of the causal explanation of why we exempt as do, going forward at least. If we do not find a justifying account then it would seem that the moral consequences of our practice of holding people responsible would decide whether or not we should abandon the practice. The actual causal story in either case would be irrelevant.

[88] Watson (1987) pg. 232

understanding requires a shared framework of values."[89] This very quickly leads Watson to the conclusion that a requirement on reasonable moral address is membership in a shared moral community, where that community is itself defined by a shared framework of values. So to hold someone responsible requires believing that they stand in a certain relationship with you, that of member of a shared moral community. There are many interesting questions of detail that could be taken up at this point, but what I want to take up is the second question that I said was suggested by the phrase 'conditions of moral address.' What is moral address? Does Watson mean to be giving us a general account of all moral utterances, or just those which are expressions of praise and blame? Does Watson even mean to be giving us an account of all blame or praise expressing utterances?

It is hard to see how he could be. After all not all cases of blame and praise involve addressing anyone at all. We can blame people without saying anything to them or demanding anything of them, whether implicitly or explicitly. I can as I write this in the early morning hours in Ithaca, NY blame George W. Bush for the second Iraq War, or Joe Lieberman for the failure of health care reform. Of course I am, in some very loose sense, in a moral community with George W. Bush and Joe Lieberman,[90] so what Watson can say is that I have the disposition to blame Bush and Lieberman and when I take myself to be blaming them I am simply aware of all the cognitive and conative elements that go into my being so disposed. This strikes me as a strained way to accommodate the intuition that we blame people privately, and it is important to notice that it relies on the fact of shared moral community, and that this fact does not always obtain when it is appropriate to blame or praise someone. I am not entirely certain that I share many moral values with Bush and Lieberman. I am

[89] Ibid. pg. 234
[90] After all, Watson says, "Obviously we do not want to make compliance with the basic [moral] demand a condition of moral understanding" Watson (1987) pg. 234

entirely sure that there would be no point to me actually telling them what moral horrors I took them to be. Whether or not they care about the prospect of their being moral horrors, they aren't going to agree with me about what goes into making someone a moral horror, and wouldn't care about my opinion anyway. So I accept that telling them what I feel about them will not change their behavior, but I don't take that to have anything to do with whether they are blameworthy or whether I can blame them.

Watson's Relationship Account is in some ways underdescribed. It is not clear how many evaluative commitments, or what kind of evaluative commitments, need to be shared between two people for it to be rational for them to blame one another. It is not altogether surprising that this is the case, because the article in which Watson presents this view is one where he is attempting only to explain one way to develop the Strawsonian position. That the proposal remains merely suggestive does not present a problem for Watson as he never goes farther by way of endorsing the view than saying the view is attractive. It would be helpful, then, to have a version of the Relationship Account that is wholeheartedly defended. Here Thomas Scanlon is of help.

4.2 Scanlon's Relationship Account of Blame

Scanlon's view is and always has been a species of the Quality of Will Account, a view that I will discuss later. That is, he accepts that people are blameworthy and praiseworthy when their actions express some morally relevant quality of will. What is distinctive about the latest presentation of his view are the conditions he attaches to something's being a morally relevant quality of will. According to Scanlon a person X is blameworthy for her action z iff[91] it is appropriate

[91] At certain points Scanlon sounds as though he is making a claim about the meaning of blame expressions, but in the interests of charity I do not think he should be so interpreted. It is obvious that he is not giving an account of what we mean when we say 'I blame X'. To see that it is enough to notice that Scanlon is giving a substantive account of blame that can be coherently denied.

for some other person Y to take the fact that X did z to show something about X's attitudes that can appropriately impair the relationship Y bears to X.[9293] To blame someone is to actually revise one's attitudes in a way that reflects the impairment of the relationship. So Scanlon's account is at once a version of the Quality of Will Account and the Relationship Account. Most importantly though, blameworthiness depends on a pre-existing relationship between the agent and the judge.[94]

While Scanlon's discussion of blame is perceptive and interesting in many ways, what concerns me is whether he provides the correct account. The problem with Watson's view is that it implausibly restricted the range of people who we could rationally blame to those with whom we shared and evaluative framework. A similar problem is going to face any relationship account. It certainly seems as though we can rationally blame anyone who has performed a wrong action, or if we can't it has nothing to do with that person's relation to us, but with her action. How does Scanlon

[92] This formula is asserted in several places. "Briefly put, my proposal is this: to claim that a person is blameworthy for an action is to claim that the actions shows something about the agent's attitudes towards others that impairs the relations that other can have with him or her. 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 to this judgment of impaired relations holds to be appropriate." Scanlon (2008)Pgs. 128-9

[93] This is the official position at least. It looks as though Scanlon at times wants to say that X is blameworthy for z when it is the case that X's doing z is itself what makes it appropriate for the relationship to be impaired. At various points Scanlon takes the fact that rival conceptions of blame cannot make sense of what he calls moral outcome luck, luck what the consequences of one's actions are, to be reason to reject the view. His official position does no better. Moral outcome luck reveals nothing of our attitudes. If Scanlon said that it was the action itself and not the attitudes it reveals which impairs the relationship then this would be closer to giving an account of blame that could handle the problem of moral luck.

[94] Scanlon does not provide an account of praise. He certainly does not think that the account he has sketched of blame works for praise. Is this a problem? I think it is, but it is not a serious problem. It seems plausible that in praising we are doing the same kind of thing as we are in blaming, but simply in a different way, or with regard to different objects. Strawson for example thinks blaming and praising are both emotional reactions people have towards actions. What differentiates them is the emotion that one has and the kind of actions to which one responds, in the case of praising the emotional reaction would be some positive or approving emotional reaction to an action one thinks is ethically good, while blaming would be some negative or disapproving emotional reaction an action one thinks is ethically bad. Praise and blame, while different, have a similar structure. Again this seems like the immediately intuitive view, one that Scanlon denies. Is Scanlon alone in this? No. Asymmetry theorists such as Wolf and Nelkin also think that praise differs significantly from blame. Given that the view that praise and blame are symmetrical is not universally intuitive apart from Scanlon, I think this is a small theoretical cost for him to pay.

deal with this problem? The obvious way to deal with the problem is to claim that the relationship in virtue of which we can blame anyone is a relationship we stand in with everyone. Scanlon makes exactly this move, citing the default moral relationship as the one in virtue of which we can blame strangers for their actions.[95] So it is because we stand in the moral relationship to everyone that we can blame anyone. This looks like it takes care of the problem of extensional inadequacy as well as can be done. It also looks like determinism couldn't be relevant to the question of whether or not we have certain relationships or whether those relationships have been impaired.

In this case, however, looks are significantly deceiving. I will present four problems with Scanlon's position. One is that it is not clear that the moral relationship is a relationship in any significant sense. I don't think I have relationships with anyone in Tibet, but according to Scanlon I do. Another problem is that even if we could be persuaded to deal with the oddness of our being in a relationship with someone we have never met, communicated with, or even learned of, there is the problem that any understanding of what it is to be in a relationship that is thin enough to cover the moral relationship that Scanlon has in mind, is an understanding on which determinism might very well imply that we fail to be in that relationship with anyone. A separate problem is that even with the positing of this odd moral relationship, Scanlon's account is still not extensionally adequate. Scanlon's view implies, despite his attempts to avoid the implication, that we cannot rationally blame the long dead for their sins. While this extensional inadequacy is not as serious as Watson's, the discussion of the issue that Scanlon provides reveals some extremely counter-intuitive implications of his view. The final and most serious problem however, is that even if there is a moral relationship, and even if we stand in that relationship everyone that we can rationally blame, Scanlon's view does not allow for moral blame at all. The basic

[95] Scanlon (2008)Pg. 138

problem is that the constituent elements of the moral relationship are such that they cannot be revised. Nor can it be appropriate to revise them. So no one is ever, according to Scanlon, morally blameworthy for what they do.

4.2.1 Is the Moral Relationship a Relationship?

Is the moral relationship an actual relationship in the ordinary sense of that word? Before being forced to answer Scanlon should ask, why does that matter? If there is only a moral relationship given some stipulated sense of 'relationship' why should this be a problem? Philosophers introduce terms of art by stipulation all the time. Even if we do not recognize a moral relationship in normal discourse, that does not mean there are not significant similarities between the moral *relationship* and actual relationships. Those similarities are presumably what Scanlon means to capture in talking about a moral relationship we have with every other person now alive. The question is whether there is an important enough similarity to justify extending the normal meaning of the term. I don't think there is. According to Scanlon relationships are "constituted by certain attitudes and dispositions. Central among these are intentions and expectations about how the parties will act toward one another. But relationships also include intentions and expectations about the feelings that the parties have for one another, and the considerations that they are disposed to respond to and see as reasons."[96]

What is conspicuously absent from this list is some kind of interaction between the people with the relationship. You cannot be friends with someone you have never met nor communicated with. In response to objections like this Scanlon offers the counter-example of the parent-child relationship. Fathers might never know of or interact with their children, but they are nonetheless fathers of those children. So, Scanlon might say, relationships do not require interaction between the people that are

[96]Scanlon (2008) Pg. 131-132

in the relationship. Does parenthood constitute an interpersonal relationship? It seems to, and very often it does, but it does not always do so. It is instructive here to consider the case of absent parents. Typically parents interact with their children, or at least know about them and think about them. In some cases this not true of course. What should we say about such cases? Suppose that someone had never met her father, and when asked about how strong her relationship with her father was she said 'I never had a relationship with my father.' Would this response be wrong or confused somehow? According to Scanlon this claim must be false, and might be a contradiction.

Of course what we ought to say here is that there are different kinds of relationships. Parenthood is a kind of relationship, but it is not the same kind of relationship as friendship, romantic love, citizenship or anything like that. The problem that this raises for Scanlon is that he defends his account of the moral relationship by reference to the properties of the one kind of relationship, the kind including friendship, and he defends the claim that the moral relationship exists by pointing to its similarities to another. But whether or not X is the parent of Y doesn't necessarily have anything to do with the attitudes or expectations of either. It is sufficient for X to be the parent of Y that X have a certain biological relationship with Y.[97] Perhaps there are accounts of the moral relationship which make it clearly of the same kind as parenthood. An account of the moral relationship on which we stand in the moral relationship to all the members of our species would do the trick. But this is not the account of the moral relationship Scanlon has offered us, and it is not clear that any relationship like friendship could possibly exist between people who had never

[97] I do not mean to deny that there is more to parenthood than a biological relationship. There are other ways to be a parent, such as adoption. The biological relationship is not a necessary condition, nor does it provide an exhaustive account of the relations that obtain between parent and child. Parents have special duties to their children. What I am claiming is that the biological relationship is a sufficient condition for the parent having those duties towards the child.

met one another.

4.2.2 Is the Moral Relationship compatible with Determinism?

Why is this seeming quibble about what kind of relationship the moral relationship is matter? Because without a clear idea of what kind of relationship is, we have no reason in particular to think that our being in this relationship is compatible with determinism. If the moral relationship were of a kind with friendship or romantic love the suggestion that its holding being incompatible with determinism would not be plausible enough to worth investigating. Friendship and romantic love are not plausibly incompatible with determinism.[98] Friendship and romantic love are relationships we stand in with other people precisely because of particular facts about them[99], and determinism, to borrow an argument from Strawson, does not imply anything of one person that it does not imply of everyone else. As long as we think it is rational to be friends and lovers with only some people and not others it is going to look quite odd to argue that determinism or indeterminism could have anything to do with. But of course what makes the moral relationship unlike friendship and romantic love is precisely the fact that we stand in the moral relationship with everyone, regardless of the properties which distinguish them from other people. So the easy argument that gets friendship and romantic love off the hook, as it were, is not available for the moral relationship.

---

[98] I do not here mean to be denying the doctrine, so important to free will theodicies, that the love God seeks from his creatures is incompatible with determinism, precisely because this is uncontroversially not romantic love. While the denial that romantic love requires indeterminism might make it harder to argue that religious love requires indeterminism, this problem can be easily dealt with by pointing out that there has been very little success providing a unified account of love. Accounts of romantic love are notoriously inadequate when applied to the familial love. As the love owed to God is often talked about as if it were owed precisely because we stand in the parent-child relationship to God, the fact that romantic love is quite unlike familial love provides reason to reject making any inferences about the love owed to God on the basis of facts about romantic love.

[99] Velleman (1999) disagrees, arguing that when we love others it is their rational agency that we love, and rational agency is something had by every rational agent in exactly the same way. This view is interesting but I have to confess I find it less plausible than just about any philosophical theory of anything that I have ever encountered.

This problem is particularly pressing for Scanlon because of what he does tell us about the moral relationship. According to Scanlon being in the moral relationship with others consists of it being the case that one ought to show a certain regard to them.[100] The basis for this duty of regard is the fact that other people "are capable of understanding and responding to reasons."[101] So if determinism is incompatible either with duties of regard holding between people, or with people being able to understand and responding to reasons, then determinism is incompatible with the moral relationship.

Both challenges have been raised of course. Kant famously argued that one could only have an obligation to do something if it was the case that one had the ability to do it. If Kant was right then if determinism obtains then no one can violate an obligation. Ishitayque Haji has more recently argued that if determinism obtains then no one can have any obligations at all.[102] Given that the moral relationship consists of default obligations we have to the people we stand in that relationship to, the moral relationship is compatible with determinism only if these arguments from Kant and Haji are unsound. That means that blame can only be appropriate, given determinism, if these arguments are unsound. What Scanlon's account of blame does, far from showing that determinism is compatible with blame, is open up a new avenue of argument that incompatibilists can take to show that blame is incompatible with determinism.

4.2.3 Scanlon and Blaming the Dead

Another, separate, worry about the moral relationship is that we do not stand in the moral relationship with everyone that we can rationally take to be blameworthy. The moral relationship requires certain kinds of regard, concern and respect from

[100] Scanlon (2008)Pg. 140
[101] Scanlon (2008) Pg. 138
[102] Haji (2002)

those in it towards everyone else in it.[103] Now it is pretty clear that we do not owe regard, concern and respect to the long dead that we owe to the living. It is also not clear that we can have the same duties to the long dead that we have to the living. So it does not appear that we can be in the moral relationship to the long dead, and neither, according to Scanlon can we blame the long dead for their actions. But this is not correct. We can rationally and often do blame the long dead for their actions. What Scanlon says about this is that blame of the long dead can only have "vicarious significance."[104] It is hard to know how to evaluative this claim. It sounds quite wrong to me. After all the Cherokee of today need not think to himself 'It was really bad for my ancestors that Andrew Jackson forced them to walk the Trail of Tears, and he did so out of racism, so it was appropriate for them to blame him.' She can just blame him, and she can do so with all the feeling and directness that that her long dead ancestors did. While Scanlon does not consider this response, or any response in fact, to his claim about blame for the long dead being restricted to blame with vicarious significance, he does have something he can say to accommodate the fact that the contemporary Cherokee can blame Andrew Jackson. Scanlon says that the relationship between perpetrator and victim is one that can stand as the ground relationship of blame.[105] Plausibly a member of the Cherokee tribe can blame Jackson for her reduced life prospects, and for the theft of land from her people. So what of the suggestion that we can blame the long dead because we are their victims, or their beneficiaries?

There are significant problems with this suggestion, any of which would doom it as a response. The first, which Scanlon himself points out but does not adequately

[103] 'Require' here does not mean 'necessary condition of' but 'obligates'. We have a default duty, according to Scanlon, to show this kind of care, respect and regard. It is compatible with our being in the moral relationship that we do not show this regard. In such a case we are failing to meet our obligations.
[104] Scanlon (2008)Pg. 146
[105] Ibid. Pg. 147

deal with, is that the relationship of having been harmed is not one that exists prior to the act for which we blame the person. In the case of the long dead and their bad acts it doesn't exist at the time the act is performed either, since we, the blamers, were not yet alive. That relationship only comes into being, presumably, at the time of our birth or sometime after. But then it is hard to see how this relationship could be the one in virtue of which we blame the dead person. After all Scanlon says that we blame people for actions when those actions show something about the quality of their will which impairs the relationship we have with them. But the actions of the long dead cannot show an impairment of an existing relationship if those actions are what create the relationship.

What Scanlon has to say about this problem has no applicability in the case of the long dead. He says that "the victim's relation to the perpetrator is impaired relative to the standard relationship between persons generally, insofar as the perpetrator's actions showed a failure to have concern for the welfare of others that is part of what we all owe to each other."[106] But I don't stand in a relationship to Andrew Jackson. I don't stand in the moral relationship to him. The moral relationship includes duties of care and respect that I can't have to the long dead, because there is no such thing as my showing the kind of respect that is morally demanded, or caring for how things turn out for them. I know of no other relationship that I could stand in to the long dead that could be the ground relationship for blame. I might be biologically related to them, but there is little that their actions could do to impair that relationship. The same goes for being in the same species as them. It might be true that being in the same ethnocultural group as a long dead person might be the kind of relationship that could be impaired. If one took being a member of the group to carry with it certain norms for behavior, the fact that a member of the group

[106]Scanlon (2008) Pg. 147

had violated them might count as a betrayal, even though the action was performed before you were born. But this suggestion, while it might work to get some blame for the long dead off the ground, does not go far enough. We can blame people who are both long dead and not members of our ethnic group. Living Jews can blame Hitler for the death of family members.

Even if there was some way of accommodating blame for the long dead into Scanlon's theory by appealing to relationships of victimhood, this relationship does not account for all the cases where one can rationally blame a long dead figure. I am of mostly German ancestry. This history of the German people was greatly affected by Charlemagne's genocide of the continental Saxons. It is likely that had this event not occurred the history of the eastern Frankish empire would have unfolded completely differently. But would that have been to my benefit or harm? Is it possible to know? Is it even sensible to ask the question? Is it even clear that I would have been born had history worked out so differently? After all my parents being where they were when they met was the result of particular historical factors (in my case the political fallout of the Soviet Union and the United States conquering the closest continuer of the eastern Frankish empire), and without those factors they might never have met. So one problem is that often we cannot actually figure out whether we have been helped or hurt by a bad action, but nonetheless we feel fully comfortable blaming the person for their bad action. Another problem is that the farther back we go in history the less likely it is that it even makes sense to ask whether the action helped or hurt me, since the action's occurring was likely a pre-condition of my existing.[107]

I know of no way that Scanlon can account for the blame of the long dead. But

[107] To clarify, this is a problem for Scanlon's account and any other account on which the permissibility of X blaming Y depends on Y having harmed X. Such views will always have trouble making sense of blaming historical figures, because the farther one goes back in history, the less likely it is that anyone in the present would exist if the action for which the historical figure is being blamed did not occur.

perhaps this is not too significant a problem. While we do blame the dead for their actions, we typically only blame the living. Watson was right to an extent; most of the time we blame people to get them to understand the wrongness of what they did with an eye towards their not doing it again. The inability to account for non-central cases of blame is a strike against Scanlon's theory, but just one, not three.

4.2.4 The Possibility of Moral Blame on Scanlon's account.

The other two strikes come from the fact that it looks like moral blame is never warranted, according to Scanlon's own theory. The basic reason is that the moral relationship is not one that it is morally permissible to alter. But the judgment that X is blameworthy just is the judgment that it is appropriate or obligatory to change one's relationship with X in light of the attitudes expressed in X's action. So if it is never appropriate to alter the moral relationship, then no one is ever morally to blame for their actions. And this is surely a significant thing to lose in one's account of blame. The constituent elements of the moral relationship that Scanlon mentions are things like intentions "to take care not to behave in ways that will harm those to whom we stand in this relation, to help them when we can easily do so, not to lie to them or mislead them, and so on."[108] These are intentions we ought to have to every other person in the world because they are our "fellow rational beings"[109] and "beings of a kind that are capable of understanding and responding to reasons."[110] So there is nothing a person could do, short of making themselves not a rational being that could make it acceptable to cease to have the constitutive intentions that make up the moral relationship. So what could blaming someone consist in then?

This is a problem that Scanlon is aware of [111] but his response is puzzling. He

[108] Scanlon (2008) Pg. 140
[109] Ibid. Pg. 140
[110] Ibid. Pg. 139
[111] "But, in contrast with the case of friendship, the basic forms of moral concern are not conditional on this kind of reciprocation. Even those who have no regard for the justifiability of their actions toward others retain their basic moral rights – they still have claims on us not to be hurt or killed, to be helped

says:

> "There is a range of interactions with others that are morally important but not owed unconditionally to everyone. If a person has no regard for the justifiability of his or her actions to others (or, despite professing such a concern, constantly sees things in a way that gives weight only to his or her own interests), then it is quite appropriate to refuse to make agreements with that person or to enter into other specific relations that involve trust and reliance."[112]
>
> and a bit later says:
>
> There is also room for modifications in our intentions to help others in certain ways. Some duties to aid are unconditional. Even murderers and rapists have a claim on us to be rescued when they are drowning or are in danger of bleeding to death after an accident. But normal moral relations also involve a general intention to help others with their projects when this can be done at little cost, and we need not have this intention toward those who have shown a complete lack of concern for the interests of others."[113]

What makes this response puzzling is that it seems to completely go back either on the general theory of blame that Scanlon has offered, or on the description of the moral relationship. If moral blame consists of refusing to trust or enter agreements or aid people in the pursuit of their projects, then the default moral relationship must contain intentions to do those things as a constitutive element.[114] But if that were the case then it would have to be the case that we ought to intend to help people in their projects, to

---

when they are in dire need, and to have us honor promises we have made to them." Ibid Pg. 142

[112] Ibid. Pg. 143

[113] Ibid. Pg. 144

[114] A potential response on Scanlon's behalf is that the duties to trust or enter into agreements with people are conditional duties. So, on this response, a duty which is in part constitutive of the moral relationship is one to honor agreements you have entered. But then what revision to the moral relationship has taken place by refusing to enter into agreements? Entering into agreements, according to this response, is not something we are obligated to do, or even be prepared to do, with other moral agents. But Scanlon's account requires that blaming constitutes a revision of the relationship. Something similar goes for trust. Suppose the duty that is in part constitutive of the moral relationship is one to trust people so long as they have not proven their untrustworthiness. How is the refusal to trust them going to count as a revision to the relationship? If the person has broken their word, and so in blaming her you cease to trust her, nothing about your obligations or intentions with regards to her have changed, at least not the obligations and intentions you have in virtue of being in the moral relationship with her.

trust them, and enter into agreements with them just because they are rational beings. And that is very implausible. I do not act immorally if I wait for people to prove themselves trustworthy before I trust them. And even if it were true, that would mean that as long as the person was a rational agent we would be obligated to have these intentions, so we couldn't abandon them. What happens if we admit the plausible thing, that trust, reliance and the like are not part of what we owe to everyone in the world just because they are fellow rational beings? Then those intentions are not part of the moral relationship, and so abandoning them would not count as a revision of the moral relationship. Put another way, Scanlon thinks that we stand in the moral relationship with others simply on account of their status as beings able to respond to reasons in the right sort of way. So, as long as they continue to be such beings, then we stand in the moral relationship to them. So how could any action on the part of a rational being make it permissible to revise or abandon the moral relationship? To do so would be to stop treating them as a rational being, and it is hard to see how that could ever be permissible, based on an action they have performed.

So there is a tension at the center of Scanlon's account. On the one hand he needs the moral relationship to be one that we stand in with every living person, and so it must depend only on those features shared by all living people. On the other hand he needs the moral relationship to be such that it can be altered. So he needs the moral relationship to both depend, and not depend, on someone's status as a person. This is a tension he faces because he wants to make blame an attitude that is about relationships. It is this commitment that I think needs to be jettisoned.

Chapter 2: The Quality of Will Account

1. Quality of Will and Incompatibilism

According to Scanlon we could blame people for their actions because of what their actions showed us about their attitudes, and because their attitudes impaired our relationship with them. We have seen why the second conjunct should be dropped, but what about the first? What about the claim that we can blame people for their actions because their actions show us something about their attitudes? Michael McKenna, and Scanlon before he advanced his relationship account, claim that remarks Strawson makes elsewhere in his famous paper, having to do with the central importance we give to having and showing good will, shows that he accepts this constraint on praise and blame.[115] As McKenna puts it, 'for Strawson, the morally reactive attitudes are responses to the quality of will expressed in a person's conduct.,"[116] and because of this Strawson accepts that questions about "being morally responsible and legitimately holding morally responsible are to be settled exclusively in terms of the moral quality of will with which the agent acts."[117] Tentatively I will say that the Quality of Will Thesis amounts to the following:

[115] I think there is good textual support for their interpretive claims. Strawson says "The central commonplace that I want to insist on is the very great importance that we attach to the attitudes and intentions towards us of other human beings, and the great extent to which our personal feelings and reactions depend upon, or involve, our beliefs about these attitudes and intentions," and that he wants to "emphasize how much we actually mind, how much it matters to us, whether the actions of other people – and particularly of some other people – reflect attitudes towards us of goodwill, affection, or esteem on one hand or contempt, indifference, or malevolence on the other." (Strawson (1974) (Pg. 5)) He also says "in general, we demand some degree of goodwill or regard on the part of those who stand in these relationships to us, though the forms we require it to take vary widely in different connections. The range and intensity of our reactive attitudes towards goodwill, its absence or its opposite vary no less widely." (Strawson (1974) (pg. 6)) When Strawson presents the division between excuses and exemptions, he is explicit that we excuse because an action does not reflect the quality of will that we thought. What he says about exemptions is less easy to construe in a way that it follows from the Quality of Will Thesis, but I follow McKenna and Scanlon in thinking this is possible.

[116] McKenna (2005) pg. 171

[117] McKenna (2005)pg. 172

QOW Thesis[118]: A person, P, is open to the reactive attitudes, blame and praise on account of some action X iff action X indicates that P has some morally significant quality of will.

So now we have another way to develop Strawson's position. It is worth noting that the Quality of Will thesis does not provide support for the Abnormality argument[119]. It may provide some support for the Actual Judgment argument. The Actual Judgment argument failed because it failed to address the possibility of generalization strategies. A compatibilist could employ the Quality of Will Thesis to block a generalization strategy. An incompatibilist might say that our particular reasons for exempting and excusing are just expressions of our commitment to a general principle like 'Do not blame people for what they can avoid'. The Quality of Will Thesis is a competitor with such incompatibilist principles, and one which does not imply anything in particular when conjoined with determinism.

At the very least the Quality of Will Thesis, if true, makes things harder for the incompatibilist. On any normal analysis of 'avoid', the 'Do not blame people for what they cannot avoid' principle looks like it, conjoined with the thesis of determinism,

[118] Scanlon (1988) never puts a principle like this forward, but I think it captures what he says about the Quality of Will Thesis in Strawson. He says "it is the nature of [the reactive] attitudes that they are reactions not simply to what happens to us or to others but rather to the attitudes toward ourselves or others which are revealed in an agent's actions." (pg. 385) This claim, when conjoined with the claim that cases of appropriately held reactive attitudes exhaust the cases of appropriate ascriptions of moral responsibility gives us the three conditions of moral responsibility that Scanlon attributes to Strawson; "It follows from this characterization that the discovery of new facts about an action or agent can lead to the modification or withdrawal of a reactive attitude in at least three ways: (a) by showing that the action was not, after all, indicative of the agent's attitude towards ourselves or others; (b) by showing that the attitude indicated in the act was not one which makes a certain reactive attitude appropriate; (c) by leading us to see the agent as someone toward whom objective, rather than participant, attitudes are appropriate." (pg. 359) Condition (c) is not clearly connected to the Quality of Will thesis, but there are ways to make it so. It has been suggested (Stern (1974), Watson (1987)) that the people covered by (c) are people who cannot take part in a moral community, and that the best way to understand this is to think of them as not being able to appreciate moral reasons. If a person is insensitive to moral reasons, or unable to understand moral reasons (where this is not the same as flouting moral demands) it seems plausible that their attitudes cannot manifest the kinds of objectionable qualities that Strawson says we are interested in. If someone really cannot understand the difference between persons and things, then their treating us like things is not so bad as someone who realizes we are deserving of better treatment, and treat us poorly anyway.

[119] I don't think there is anything that can help the Abnormality argument.

implies that we are not blamable for what we do. There is no such obvious implication with the Quality of Will Thesis. But that does not mean that there is no such implication. The incompatibilist can press two points. The first starts with the question, 'Are we responsible for the quality of our wills?' If the **Quality of Will Thesis** gives us necessary and sufficient conditions for moral responsibility, we cannot be responsible or for qualities of our will, but only actions. But the incompatibilist will say that they have asked a sensible question, and so that it is not just a category mistake to ask whether we are responsible for the qualities of our will. If it is not a category mistake, then we can be responsible for the qualities of our will or we can fail to be responsible for the moral qualities of our will, and the incompatibilist can safely appeal to the following principle; if X is explanatorily sufficient for Y and we are responsible for Y, then we are responsible for X[120]. The best explanation of why it is that actions express or indicate the quality of one's will is that one's quality or qualities of will explain the action. So, if the incompatibilist can fairly appeal to a transfer of non-responsibility principle, the Quality of Will Thesis is left only with the implausible theory that each quality of will is one we are responsible for, which means that it indicates some other quality of will. The only way, it seems to me to avoid this conclusion is to argue either that the incompatibilist cannot appeal to a transfer of non-responsibility principle, or argue that it is, contrary to how it seems, some kind of category error to ask of a quality of our will whether we are responsible for it.

The second point that the incompatibilist is likely to press is the more interesting of the two, and the more likely to express a deep problem. Some incompatibilists will claim that for our wills to have a moral quality at all, that quality

[120] This is equivalent to the more common transfer of non-responsibility principle, according to which if X→Y and some person P is not responsible for X then P is not responsible for Y. Both my transfer of responsibility (which goes backwards along lines of causal sufficiency) and transfer of non-responsibility (which goes forward) principles are of course rough versions that any incompatibilist

must be one we autonomously chose to have, or at least it must be a reflection of our autonomy. They will claim that autonomy is central to morality, that it is our autonomy that makes us moral agents, and that no morally praiseworthy quality can hold of us if we are not autonomous. With these claims in hand they will then go on to claim that autonomy requires freedom in a strong enough sense that it is incompatible with determinism. The claim here is not that the Quality of Will thesis is false, but that appealing to it to settle the issue between compatibilists and incompatibilists misses the point, because any plausible account of moral worth brings along with it commitments that are in tension with determinism.

So there are two ways in which incompatibilist can respond to the Quality of Will thesis, on the one hand by questioning whether qualities of will can be the fundamental facts about responsibility at all, and the other by claiming that the Quality of Will thesis, when its presuppositions are laid bare, is potentially an incompatibilist principle. Now I do not think any account of blaming and praising is going to give compatibilists the resources to rebut the criticism that to have a will with moral qualities we require the kind of freedom that is incompatible with determinism. But even if this route to the incompatibilism is not ruled out by the Quality of Will thesis, if it could rule out all the other ones, then this would be a significant achievement. If it could answer the problem of the recursiveness of moral responsibility, along with show why it is that the ability to do otherwise and being an ultimate source of one's actions are irrelevant to whether or not a person is morally responsible for her actions, then the view would be very significant. So it is worth looking into whether the Quality of Will thesis, or something like it, is true.

On the Quality of Will thesis to blame someone is just to make a judgment about their will, and to blame someone is justified if and only if the person knows that

the judgment is true.[121] What kind of judgments are these? I take it that the paradigm judgments are one's like 'He acted cowardly' or 'He acted in a very caring way.' Such judgments explicitly relate actions to psychological features the agent might have.[122] But blaming judgments could take different forms as well. 'That was rude' is an assertion that is technically about an action, but it identifies a feature of an action that typically is explained by someone having a certain quality of will. There are some arguments against the thesis that blaming consists entirely in a judgment that strike me as odd. The first is the association of this thesis with the 'moral ledger' account of moral responsibility. This account, put forward by Jonathan Glover[123] says that when we blame someone we are doing something like record keeping.[124] To blame someone is to hold something against them by marking it down in their moral ledger. This account is very odd, and almost no one who writes on praise and blame accepts it. It suggests that what we care about in blaming and praising is not the rightness or wrongness of the action for which we are blaming and praising the person, but how this affects the overall judgment of that person's life. And not only is it not true that we care about such an overall evaluation more than the evaluation of the person for a particular action, it is not clear we care about the overall evaluation at all. This is not to deny that we care about the character of a person. We do. But the moral ledger is not the same thing as character. A person with a long history of lying can become honest, for example.

The moral ledger account has little to say for it, but there is no reason to think

[121] Later I will present a version of the Quality of Will account that does not have this implication about the justification of praise and blame. It is that version, and not the one currently under consideration, that I endorse.

[122] Blaming judgments need not mention stable psychological features of the agent. Even if a person is normally courageous we can, with sufficient evidence, reasonably conclude on some occasion that he acted out of fear that he ought to have resisted.

[123] Glover (1970)

[124] Michael Zimmerman (1988) also endorses a view like this, but he is wary enough of the 'moral ledger' metaphor that it is worth taking this attribution with a grain of salt.

that the thesis that blaming is constituted by a judgment implies the moral ledger account. The moral ledger account goes beyond the thesis that blaming consists in making a judgment and gives an account of what type of judgment we are making and why we should care about that judgment. The Quality of Will thesis need make no such particular commitments simply in affirming that blaming consists in a judgment.[125] Another point worth making is that the Quality of Will thesis need not be committed to the claim that blaming does not involve an emotional aspect.[126] As we saw in the discussion of the Emotion View, there are some ways of thinking about emotions on which they include judgments, so that one could make a judgment about someone's quality of will by being resentful, for example.

So some common reasons to reject the Quality of Will thesis fail. Does the Quality of Will thesis rule out traditional incompatibilism? It has been argued, by Nomy Arpaly, that on a certain interpretation of incompatibilism it does. I said earlier that the best way to understand the incompatibilist position was as a moral position, specifically a position concerning how it was fair to treat people given certain facts about how the world works. There are more reasons to accept that incompatibilism is a doctrine about how it is fair to treat people than just the usefulness of this interpretation for responding to Strawson. One good reason to accept incompatibilist principles like the principle of alternative possibilities or the principle of transfer of

[125] So I do not think that the Quality of Will account can be fairly criticized as reducing blame and praise to 'mere grading' as Scanlon (2008) and Sher (2006) do. 'Mere grading' suggests a shallowness and pointlessness that need not characterize praise and blame on the Quality of Will account. Glover's account of responsibility, which can fairly be thought of as one version of the Quality of Will account might have been guilty of making blaming and praising out to be shallow pointless activities, but that was because of the way he constrained the type of judgments that constitute blame and praise. After all how important is it to know that someone did so and so many bad things versus so and so many good things? If the range of judgments is expanded then the view does not make blame and praise shallow or pointless. Then instead of just keeping count of instances of good or bad behavior we can say that blame and praise is a matter of making judgments like 'You did that out of pure spite', or 'She was animated in her sacrifice out nobility just as much as defiance.' The depth and point come from the judgments that are licensed, not the general account of blame and praise.

[126] Passé George Sher (2006).

non-responsibility, or a control principle of moral responsibility, is that these principles reflect basic norms governing fair treatment of persons. So a reason to accept the Principle of Alternative Possibilities is that it seems to follow from an intuitive principle which says that it is not fair to treat someone differently on account of something they could not avoid. In addition, as I said before, this way of interpreting incompatibilism removes any air of over-intellectualization from that position.

2. Arpaly's Challenge to Fairness Based Incompatibilism

It might seem that if the Quality of Will thesis is correct then incompatibilism so interpreted just has to be false. As Arpaly says:

> "Punishing a person is an action: it is something one *does*. Blaming a person is not an action – I might blame Brutus for the death of Caesar without doing anything – though the verb *to blame* can also be used to refer to the act of notifying someone that you blame a person. The observation is hardly new, yet philosophers tend to dismiss the difference between blaming and punishing or remonstrating; we discuss whether it is fair to blame a person as if the question were equivalent to the question of whether it is fair to punish the person. This is a category mistake. The primary sense in which I can be fair or unfair in blaming someone is the sense in which believing that Ron is an idiot might be fair if Ron is an idiot and unfair if Ron is not. The primary sense in which I can be fair or unfair in punishing someone is the sense in which my calling Ron an idiot might be fair if he has just called me a moron and unfair if he has never been rude to me."[127]

There are two points made here, one of them used to argue for the other. First Arpaly says that blaming is not an action, and, I take it, argues from this claim to the fact that different standards of fairness are at issue in blaming than are at issue in punishing. I object to both points. I do not think that making the judgments involved in blaming and praising differ significantly from action, and so I think it can sometimes be unfair to believe something that is true and that you would be epistemically justified in

[127] Arpaly (2006) pg. 9

believing is true. So to continue Arpaly's example, let's say all the evidence is in support of Ron being an idiot, the evidence is substantial, and there is no reason to think that it is pointing you in the wrong direction because of peculiarities about this case. It still might be unfair to Ron to think that he is an idiot. Do we really think it is fair to think of Ron as an idiot when the reason Ron is an idiot is because his mother drank herself silly every night she was pregnant? I don't think so. I think it would be wrong of us to think to ourselves 'Ron is a real idiot'.

2.1 Judgments and Actions

Arpaly's first point, that blaming is not an action, is just wrong I think. Certainly blaming is not necessarily a public action, and Arpaly is right to insist that notifying someone that we blame them is not the same thing as blaming them. It is compatible with this that blaming is a private action. What about the example of blaming Brutus for the murder of Caesar? What action is being performed in blaming Brutus? I agree with Arpaly that it is not the act of exclaiming 'Damn you Brutus!' or anything like that. I suggest that the action is that of coming to believe something along the lines of 'Brutus wrongfully killed Caesar,' or deciding that Brutus wrongfully killed Caesar, or assenting to the proposition that Brutus wrongfully killed Caesar.[128]

It might be thought that this is not an action, but I see no reason to think that, especially given that we are wondering whether it is an action for the purpose of deciding what standards of fairness attach to it. After all Jane can ask Tom to not think ill of her for going to Europe for the summer instead of spending the last months before college with him, and if Tom does promise this and goes on to think ill of her, Tom is being unfair to Jane. He is being unfair because he promised not to do that.

[128] While I attach no importance to these different ways of describing the action, I can understand that some might, and so I do distinguish between these different descriptions later.

So believing something or thinking in certain ways can be unfair, and can be the kind of thing we promise to do. What reason is there to deny that they are actions?[129]

2.1.1. Doxastic Voluntarism

The most popular reason given as to why coming to believe something is not an action has to do with the voluntariness of beliefs. Actions are directly under our control, and beliefs are not, the argument would go.[130] The claim that beliefs are not under our control or are less under our control than our actions are is actually a problem for fairness based incompatibilism in a more direct way than just by suggesting that beliefs are not actions. If our beliefs are not under our control, then even if coming to believe something or making a judgment might count as an action, it looks as though it is not the kind of thing we could be obligated to either do or refrain from doing. Why? To see consider two claims:

> **Doxastic Involuntarism**: No one has the ability to believe X when they also believe that there is conclusive evidence that X is false.
> **Ought Implies Can**: If X ought to Y, then X has the ability to Y.

What these two principles together entail is that no one can have an obligation to believe something that they believe there to be conclusive evidence against. I take the argument from **Doxastic Involuntarism** and **Ought Implies Can** to be unsound, because I think **Ought Implies Can** is false. More importantly though the implication of this argument is perfectly acceptable to me, as my case that praise and blame can sometimes be unfair does not require that we have obligations to believe something

[129] There is a point to be made here that one can blame Brutus long after one has decided that Brutus wrongfully killed Caesar. Once one decides that Brutus wrongfully killed Caesar one counts as blaming Brutus until one's mind gets changed. But actions are not similarly temporally extended. To get around this problem let me say that there is the act of blaming that is constituted by a person coming to believe or making a judgment, and a state of blaming which is the product of the action of blaming. Presumably the standards of fairness that govern the action of blaming also govern the state of blaming. Tom can't defend himself to Jane in September by saying 'I think ill of you now, but I decided you were mean a long time ago, why are you still mad at me?'

[130] Defenders of Doxastic Involuntarism include Williams (1973), Feldman (2000), Dretske (2000), Pojman (1993), Alston (1998), Plantinga (1993), Scott-Kakures (1994), Curley (1975)

one believes the evidence shows conclusively is false.

It is not because I think that while reasons of fairness sometimes speak against making certain judgments there are no duties to refrain from making those judgments. There being reasons of fairness against X amounts to there being a pro tanto duty to refrain from X-ing, and so in some circumstances there will be all things considered duties to not make certain judgments. So there are obligations related to beliefs. This does not imply that we are obligated to believe something we believe we have good evidence is false. It is important to see that one can defend fairness-based incompatibilism against Arpaly's attack simply by establishing that it is sometimes unfair to make certain judgments about people that are well supported by the evidence. This conclusion is not threatened by the argument from **Doxastic Involuntarism** and **Ought implies Can** for two reasons. One is that it does not require that people believe anything, and the other is that it does not require anything about cases where the evidence conclusively establishes the truth of some proposition. Incompatibilists do not claim that if determinism is true then we ought to praise rather than blame, or vice versa. Incompatibilists claim that we should do neither, and according to the **Quality of Will Account** refraining from praising and blaming is the same as refraining from making certain kinds of judgments. The judgments in question are judgments concerning the quality of another person's will, and such judgments are always made under uncertainty. Often a person's own intentions, desires and values are not clear even to them. They are usually less clear to others. So when we blame and praise people, according to the **Quality of Will Account** we are making decisions that are strictly speaking underdetermined by the evidence.

What is clear is that we need some stronger version of **Doxastic Involuntarism**. So let me suggest:

> **Strong Doxastic Involuntarism**: No one has the ability to believe something she takes to be not best supported by the evidence and no one has the ability to refrain from believing something she takes to be well-supported by the evidence.

Strong Doxastic Involuntarism does, when conjoined with **Ought implies Can** entail that there can be no obligation to refrain from making a judgment that is best supported by the evidence. This would mean that considerations of fairness could not apply to praise and blame beyond considerations of evidential fairness. So if **Strong Doxastic Involuntarism** and **Ought implies Can** are both true, then the **Quality of Will Account** is not neutral between compatibilism and incompatibilism.

I have already said that I think that **Ought implies Can** is false, so why don't I just deny the soundness of the argument by denying **Ought implies Can**? There are two reasons. The first is that **Ought implies Can** is a widely accepted and the case against it relies on claims in ethics and meta-ethics which are, if anything, generally thought less plausible even than the denial of **Ought implies Can**. The second reason is that even though I think **Ought implies Can** is false, there is a similar principle which is almost certainly true and would do all the work in the argument that **Ought implies Can** does. In brief, **Ought implies Can** would be true if the correct analysis of ability was some kind of conditional analysis of ability. Unfortunately the conditional analysis of ability is false[131], but this should not stop us from appreciating

[131] As a universal account of what we *mean* by the word 'can' or the word 'ability' the conditional analysis is a failure. There are just some occasions where people utter those words and do not mean what the classical compatibilist supporters of the conditional analysis said they do. This does not automatically show that the conditional analysis is not the proper analysis of the words 'can' and 'ability' in contexts in which principles like **Ought implies Can** is uttered. In certain contexts the conditional analysis looks like the appropriate one. When asking whether a bridge can support a large truck, what we are asking is whether, if you put the truck on the bridge the bridge would stay up. I say confidently that the conditional analysis is false because it looks like in contexts where moral questions are at issue the conditional analysis is the wrong analysis. Keith Lehrer (1968) has provided examples in which the conditional analysis gives the wrong answer to the question, 'Could he have done X', and these examples are of just the kinds of cases that would raise moral issues. Is it possible that we can distinguish contexts in a more fine-grained way, such that there is one sense of 'can' in cases of **Ought implies Can** which is not operative in other moral contexts? We could, but it is hard to see how such a parsing of contexts and senses could amount to an argument for a compatibilist position. At most it

that everyone who is part of the free will debate seems to accept the following principle:

> **Ought implies Conditional Ability**: If X ought to Y, then it must be the case that if X φ-ed Y-ing, then X would have Y-ed.

In this definition φ is a placeholder for some kind of mental item, be it an event or state. Different versions of the conditional analysis of ability substitute different mental items for φ. Some substitute 'choosing', while some substitute 'intending to' and others substitute 'have one's strongest desire in favor of' for φ. The details are very important if one is interested in defending a conditional analysis of ability. I am not. As an account of the meaning of the English words 'can' and 'ability' I think the conditional analysis is hopeless. More plausible is a view on which the truth of some conditional is necessary and sufficient for a person to have the ability in question, but on which no claims about meaning are made. But even such a conditional account of ability faces problems, some of which I will talk about at the end of chapter four.

But for present purposes, it is not necessary to defend either a conditional analysis of 'ability' or a conditional account of the nature of ability. It is just enough to point out that almost everyone is going to agree to **Ought implies Conditional Ability** and that taken together with a suitably revised version of **Strong Doxastic Involuntarism**, this weaker principle entails that it cannot be the case that we are obligated to not make judgments that are well supported by the evidence. So for fairness based incompatibilism to be consistent with the **Quality of Will Account**, **Strong Doxastic Involuntarism** must be false.

I will argue that **Strong Doxastic Involuntarism** is false. **Strong Doxastic Involuntarism** says that we cannot refrain from making judgments that we think are well supported by the evidence. But it is relatively obvious that this is false. We

---

could show how a compatibilist could talking about free will without giving up on her commitment to any intuitively plausible principle.

refrain from making the judgments we take to be best supported by the evidence all the time. There are more intellectual vices than just ignorance. Intellectual laziness, intellectual cowardice and other forms of lacking intellectual integrity all lead to people not drawing conclusions that they know are well supported by the evidence. If I am presented with compelling evidence that my friend has betrayed me, we do not need to attribute to me ignorance of the facts or some kind of logical error, if I refuse to make the judgment that my friend betrayed me. Perhaps I found the thought so loathsome I simply start thinking of something else entirely, like what I need to get at the store today. Perhaps I grew angry and immediately took my anger out on something that has nothing to do with issue of my friend betrayal, like the poor functioning of my phone. If the doctrine of doxastic involuntarism forces us to deny that such events occur, then the doctrine is simply false.

The defender of the doctrine might rightly point out here that she makes exceptions for self-deception, but wrongly be tempted to say that there is self-deception going on here, that in refusing to draw the conclusion I know to be warranted I am fooling myself or tricking myself. But this is implausible. How can I be deceived about the issue if all my beliefs about the issue are true? I haven't made a judgment that I might have made, that would have been true, but I haven't deceived myself at any point. So I think it is obvious that whether or not we make certain judgments can be determined by our will, in that we can simply decide to do something besides finish thinking the issue all the way through.

But is this 'stopping short' move the kind of connection between the will and judgment that I need? Can it be used by someone who has good evidence that X to not make the judgment that X, because X is unfair? It seems not, because in thinking about the unfairness of judging that X, one is still basically engaging in the same activity, that of considering whether one ought to judge that X. Let us say Jack has

been presented with evidence that his wife is unfaithful. Jack sees the evidence and knows it is substantial, but also knows that he owes his wife some trust and faith, because he promised to give her that trust and faith. So he knows that it is unfair to judge that she is committing adultery.[132] Perhaps at the moment he realizes that the evidence points towards the fact that she has committed adultery he thinks to himself 'It would be unfair of me to think that' and to stop himself from thinking it, immediately starts thinking about something else. This might work, but it is odd to say that Jack could think about the proposition 'My wife has cheated on me' well enough to realize it is unfair, but not well enough to come to believe it, given that the evidence is right in front of him. It seems that a more natural way to describe what is going on is to say that Jack comes to believe that his wife cheated on him, and then represses that belief, or disavows it right afterward.[133]

In any case, to assimilate all cases of refraining from making a judgment to cases of 'stopping short' would be incorrect. We do not always refrain by means of this kind of manipulation of what has our attention. Sometimes we consider a thought and refuse to endorse it, or commit ourselves to it as true. Think of Descartes in the first several Meditations. A wide range of beliefs occur to him, and occur to him naturally as the explanations of the experiences he is having, but he refuses to assent to them. In Descartes' case he refrains because he has accepted the principle that any proposition which might be false is one that cannot serve as the foundation of

[132] One might think that if the evidence suggests she has committed adultery, and thus broken her own promises, that Jack is no longer bound by his own promises. That might be true. But I take it that the promises Jack made ought to enter into the argument earlier. Given that Jack promised trust and faith, he ought not think 'my wife is an adulterer' without the evidence being overwhelming, so as to lead to near certainty. This standard of evidence is not the same, I take it, as the standards one ought to use in normal belief formation. So let me add to the case that the evidence before Jack is good enough to meet the standards for normal belief justification, but fall short of the near certainty required for him to believe that his wife is an adulterer without at the same time treating her unfairly.

[133] It might be that the demands of fairness are such that this would be enough to count as Jack treating his wife fairly. I do not think that making the judgment and then disavowing it would work for the issue of fairness in praising and blaming.

knowledge. Descartes refrains from judgment for theoretical or epistemic reasons, and so it might be that his case can be accommodated by **Strong Doxastic Involuntarism**. But, in this case a thought reliably occurs to Descartes, and occurs to him as the proposition best supported by the evidence, and it is only on reflection that he comes to reject the proposition, not because it is not best supported by the evidence but because it is not well supported enough by the evidence. There is a difference between a thought occurring to one as correct, and assenting to the proposition, or making the corresponding judgment.

I think that **Strong Doxastic Involuntarism** is false because it treats everything that can fairly be described as 'coming to have a belief' as similar to the occurrence of a thought that seems correct. Plausibly we do not have any control over the occurrence of such thoughts. But that does not mean that we have no control over the attitude we take to such thoughts. We can make the judgment that the propositions are correct, or we can refrain from doing so. And it seems to me that my control over doing so is just the same as my control over whether I pour a glass of juice in the morning. I think we are all aware of cases where we refrain from making the judgment that seems right on the basis of theoretical reasons, and as I said this is not a case that need trouble the strong doxastic involuntarist. When we refrain from assenting to a proposition because we have adopted higher than normal standards of evidence, even though our judgment is not determined entirely by our judgment about the balance of evidence, it is still being determined by the evidence in a way, because it is being determined by principle about the evidence, in this case a very rigorous one. The question is whether we ever refrain from assenting to a proposition for non-theoretical reasons.

That people do this is a settled fact of common sense, but of course that doesn't make it obviously true. Arguments very often end with one or both parties

complaining that the other is not being reasonable and just believing what he wants to believe.[134] People resist conclusions of arguments for which they have no adequate response when the conclusion threatens a belief around which they have structured their life, or which makes their life seem significant or meaningful. If distaste with a belief, or a desire to continue living as one has can led a person to refrain from making a judgment in accordance with the evidence, why can't moral principles do the same? Why couldn't principles of fairness lead people to refrain from making judgments about the quality of another person's will? And if it would be unfair to make a certain judgment, then why wouldn't it be the case that we ought to refrain from making that judgment? Is there ever, in cases of decision under uncertainty, a comparable epistemic obligation to make a judgment? I doubt it. So, the hypothesis is the making of judgments of certain kinds is subject to standards of fairness.

2.1.2 Doxastic Transparency

One fact Arpaly might point to is that we do not always know what our beliefs are, and it is commonly thought that we are always aware of our own actions. In fact, some have claimed that it is an essential feature of actions that we have direct unmediated knowledge of them, and that is certainly not true of beliefs.[135] Our beliefs are often a mystery to us. Often when we reconstruct our own reasoning about a problem we find that we cannot understand it without appealing to beliefs we did not at the time know we had. We can have unconscious beliefs of course, and while we can come to know what those are, usually doing so requires the help of a professional. Perhaps the easiest way to appreciate this supposed difference is to think about

[134] Such an accusation is actually stronger than anything I need to argue for, as I simply claim that the ability to refrain from making a judgment is in our power. In some cases the ability to refrain might suffice for being able to believe what one wants. Suppose one comes to believe something as the result of evidence and later on is presented with more evidence, so that the balance of evidence now supports changing one's mind. I think that people have the ability to refrain from doing so, thus leaving them with the original belief, with which presumably they are more comfortable.

[135] Velleman (1989)

dispositional beliefs. Dispositional beliefs are beliefs which we have, but which are not currently occurring to us. But even when I am not thinking about black holes, as I usually am not, I still have the belief that for every black hole there is a singularity. I know this because of how unsurprising it is for me when, after having been prompted for information about black holes, I say something like 'Well every black hole has a singularity.' I did not just discover the belief all over again, I simply recalled it from where it was being kept. But sometimes, when we have not had to recall a belief for a while, it can take some time to recall what it is. We might think to ourselves, 'Where did I come down on the question in the end?' Sometime we even try to remember the very last occasion on which we confidently announced our view on the issue, so as to remember which view is ours. This is not a form of direct unmediated knowledge, this is knowledge by inference.

This would be a significant worry if it were true that all actions are such that we have direct unmediated knowledge of them, but there is reason to be suspicious of this claim. Some of the same problems that faced beliefs can be brought up with regards to actions. We are not, for example, conscious of all our actions. The well worn example of the driver not realizing he had already taken his exit is well worn because it is a good one. The driver turns the wheel exactly as much as he should in order to take exactly the lane he needs to in order to get where he is going, so this is no mere twitch. It is also something the driver is not aware of until after it happens and he comes to know it happened by inference. He looks around, realizes he is on a different highway than the last time he checked, and figures out that he must have taken the exit without being aware of it. And what he is aware of is the fact that he took the correct exit, that is, he is aware of an action he performed. If he took the exit because it is normal route home, but today he was making a delivery to somewhere that required his taking a different exit, his failure can fairly be attributed to him,

rather than his muscles, as with a twitch, some external force, as with non-actions like falling down, or to bad luck, as with accidental behaviors such as stepping on someone's foot. Now defenders of the claim that we have direct unmediated knowledge of our actions have responses to criticisms of this kind. Perhaps the claim will be restricted from 'All actions are such that we have direct unmediated knowledge of them' to 'All non-habitual actions are such that we have direct unmediated knowledge of them'.

Such a revision would take care of the driving example, and might provide us with a true claim. The question is whether this claim is any longer of interest. The best argument, based on this claim, that someone like Arpaly could give is that beliefs are not non-habitual actions. And one might wonder whether the implications of the classificatory claim are interesting. It was certainly interesting to be told that blaming was not an action because to blame someone is just to believe certain things about them. That something is an action is of immediate moral importance. Certain standards apply to actions that do not apply to other kinds of events. But is there any special significance to 'non-habitual action' over and above the significance of 'action'? I do not think so.

And even if there were there is every reason to meet the revision from 'action' to 'non-habitual action' with some similar revision about the kind of belief we are talking about. Arpaly claims that blaming cannot be an action because it is a kind of belief. And perhaps not all beliefs are actions, but that does not mean some aren't. Perhaps having an occurrent belief is an action. It certainly seems to be something we have direct access to in the same way we do our own actions.

3. Judgments and Fairness

So the argument from blame's status as a non-action to the different standards of fairness for blame and punishment is unsound. We have duties of fairness

concerning private mental events like blaming, just as we do concerning public actions like punishing. But this does not mean that the same considerations can count in favor of or against blaming as count in favor of or against punishment. Whether we punish people, and the extent to which we punish them is determined in part by matters of social policy. Truth and Reconciliation tribunals will often withhold punishment of those who deserve blame because it is the only way to achieve certain social goods. We punish negligence in ways that far outstrip worthiness of blame; if Jack slips on Oliver's sidewalk because Oliver has not taken the time to shovel the snow off it, Oliver is open to sanctions that outstrip his blameworthiness. Oliver did not intend to make Jack slip, did not think, let us suppose, that by not shoveling someone would be made to slip, and did not fail to think things through. I am imagining that Oliver thought to himself, no one will be out in such a storm, and Oliver thought this because Oliver could not imagine going out in the storm himself. I think Oliver is morally to blame to some extent. The thought that it was possible that someone would be out in that storm ought to have occurred to him, and he ought to have, in reaction to that realization, shoveled the snow just in case. But in the context of punishment there is a burden that has to be put on someone, the burden of paying for Jack's medical care and making up for the time Jack has lost. And it seems clear that burden should fall on Oliver, regardless of how blameworthy Oliver is.

So punishment and blame are not subject to all the same standards. But are they as different as Arpaly thinks them to be? I think they are not. Recall that according to Arpaly, "The primary sense in which I can be fair or unfair in blaming someone is the sense in which believing that Ron is an idiot might be fair if Ron is an idiot and unfair if Ron is not."[136] I have two criticisms to make, one which I think is obviously correct, and the other which is less than obviously correct. Luckily the

[136] Arpaly (2006) pg. 9

obviously correct criticism can be used to prop up its weaker partner. To start with it is certainly not fair to Ron to believe he is an idiot just because he is an idiot. One might have no reason to think Ron is an idiot. Perhaps you are Ron's cousin and only see him on the holidays, and during those gatherings he never acts like an idiot. One would be treating Ron unfairly to think that is he is an idiot, because one's belief that he is an idiot could not have been reached by a fair evaluation of the evidence. As I will put it there are reasons of evidentiary fairness that make it unfair to believe that Ron is an idiot in this case.

How far do reasons of evidentiary fairness go towards restricting the judgments it is fair to make? They cover cases where the evidence at your disposal shows that the belief in question in false. They should also rule out making the judgment that X is a coward when you know that there might be information which shows that X is not a coward, and that you could rather easily gain access to that information. Similarly obvious is that it is unfair to judge that X is a coward without considering certain forms of evidence that are always relevant to judgments of cowardice. It would be unfair, for example, to judge that soldier was a coward for retreating without knowing whether he had been ordered to retreat, or without knowing what the costs of not retreating are.

There are, however, less obvious ways to violate the duty to evidentiary fairness. Even if one has, relative to a particular situation, all the evidence that one can be reasonably expected to gather, and all that evidence supports a particular judgment, if you also know that there are some regular ways in which evidence fails in cases of this sort, evidentiary fairness would demand that one withhold judgment. So, for example, if all the eyewitness testimony in a criminal case supports judgment X, but the judge and jury know that eyewitness testimony tends to be unreliable in cases of this sort, then rules of evidentiary fairness would demand that we do not endorse X,

even though all the available evidence is in favor of X.

So it is not the case that it is fair to belief p just because p is true, nor is it fair to believe p just because one has more evidence for p than against it, nor even is it true that it is fair to believe p just because one has more evidence for p than against it and that evidence is normally sufficient for justification. And this all from principles of fairness that concern how to deal with evidence. But what seems pretty clear to me is that there are principles of fairness governing belief that go beyond principles of evidentiary fairness.

I think it is clear that historical considerations can alter what judgments are fair to make about a person. In Ron's case I have already said that I think it is unfair to Ron to think of him as stupid if his stupidity is the result of prenatal neglect. Something similar goes for the athletic adult who sees competitors in the Special Olympics and says to themselves 'I am so much more athletic than them.'[137] This could certainly be true. Nonetheless it is just wrong to think that. One should not think that. And, I think, one shouldn't think it not because it isn't true, but because to evaluate those athletes in this way is unfair to them. Ron's idiocy and the Special Olympics do not involve moral judgments of the kind that Arpaly thinks are involved in blame.

To remedy this problem let us consider Aunt Harriet. Aunt Harriet has had it hard. She was married early, against her will, because of pregnancy. After she lost the baby she stayed married, against her will, because of Catholicism. Her husband neglected her and every few months abused her until the day he died with huge debt hanging over the two of them, which Aunt Harriet had to spend the rest of her life working off. Now in her 70s she is terribly rude to everyone. She never sets out to

[137] Distinguish this judgment from one that says 'I am so much more impressive than they are, because I am more athletic.' This judgment simply isn't true. Competitors in the Special Olympics have faced and overcome obstacles that are far more significant than the extra minute they might take to run the mile.

hurt people, but she says very rude things and fails to be considerate of others' feelings. Is Harriet unkind? Yes. Is she a rude person? Yes. Is it fair to think to yourself 'Aunt Harriet is rude hateful woman'? No, it isn't fair. She is rude because of mistreatment. Blaming her for her rudeness is unfair. This is not to say that you should not take the evidence for her being vicious and unkind into account when planning, for example, whether to vacation with that side of the family with your sensitive young son. The judgment that she is likely to hurt his feelings is distinct from the judgment involved in blaming her.

4. Praising and Blaming as Benefiting and Harming

What could explain its being the case that reasons of fairness tell us that we ought not to make certain judgments, even when those judgments are just in our head? In order to answer this question I will have to argue for one of the central claims of my dissertation, that, on the Quality of Will account of praise and blame, to praise or blame a person is to benefit or harm them.

> **PBBH:** If X praises Y, X benefits or does good for Y, and if X blames Y then X harms Y.

This claim is central to two of the broad goals of the dissertation, that of providing a conception of moral responsibility that is neutral between compatibilism and incompatibilism about determinism and free will and moral responsibility, and that there is an argument for incompatibilism that does not rely on contested concepts like control or ability. How claiming that praising and blaming constitutes benefiting and harming would help establish a neutral conception of moral responsibility is hopefully by now clear.

If the Quality of Will account of praising and blaming is correct, then praising and blaming involves, centrally, the making of judgments about the agent being praised or blamed. As we have seen the Quality of Will account avoids the explicit

compatibilist commitments of the Relationship account and the Emotion account of praise and blame. Arpaly has argued, however, that we are warranted in making the judgments involved in praising and blaming just in case those judgments are true. This leaves no room, as Arpaly points out, for considerations of fairness to play a role in determining whether we are warranted in praising or blaming people, because considerations of fairness do not have anything to do with whether the kinds of judgments involved in praise and blame are true or false.

This result is of significant concern for me, both because it undercuts my argument against the Quasi-Realist interpretation of the Emotion account, and because it would mean that the Quality of Will Account was ruled out by traditional incompatibilism.[138] I have argued, by appeal to examples, that Arpaly is wrong and that considerations of fairness are relevant to the warrantability of the judgments involved in praising and blaming. While I am entirely satisfied by that argument I anticipate that others will not be content with an argument that relies so heavily on suggestive examples, so I wish to present an argument from general principles. The most controversial of those principles is that praising and blaming are instances of benefiting and harming. It is for this reason that standard uncontroversial principles of fairness apply to making the moral judgments involved in praising and blaming, and why it is that the warrantability of those judgments depend not just on the truth of the judgments or even the evidence available in favor of the judgments but also the fairness of making the judgments.

The argument that praising and blaming can be unfair appeals to two principles, one of which is controversial, the other of which is nearly obvious. The controversial principles is **PBBH**. The nearly obvious principle is that it is unfair to

[138] It would still, as I said, be compatible with versions of incompatibilism which claim that determinism is incompatible with moral rightness or wrongness or that claim that determinism is incompatible with anyone having a morally worthy or unworthy will.

harm or benefit someone that does not deserve to be harmed or benefited.

4.1 Benefits, Harms and Fairness

While I think that the second principle is nearly obvious I have to admit that its entailments sound a bit odd. It entails that when someone receives a benefit that they did not deserve, that is unfair. This sounds harsh. Don't we praise, and by hypothesis benefit, children for things they do not deserve to be rewarded for? Don't we give people rewards all the time for things like winning the lottery or at gambling? We certainly do those things. The question is whether it is fair to do so. What is odd about the principle is that it says that it is indeed unfair to benefit people in this way. Why should there be limits to how we ought to help people? Isn't it always right to help people? And how could it be right to do what is unfair? Another problem is that it is hard to see who we are being unfair to in cases of undeserved benefit. It is hard to see how we could be acting unfairly towards the beneficiary. He has no reason to complain, and being treated unfairly would provide him a reason to complain. Perhaps in many cases of undeserved benefit there is someone who did deserve that benefit but who did not receive it because some undeserving person did. In that case the deserving person is being treated unfairly. But it need not always be the case that there is some more deserving person being neglected when we benefit someone who does not deserve it. And even when there is this seems as though it only matters when the benefit is scarce. If the benefit is plentiful then to give it to the undeserving is not to neglect giving it to the deserving. And praise is not a scarce resource, but one we can create anytime we need it.

I turn first to the problem of our praising children. The problem is supposed to be that we praise children even though they cannot deserve it. And in most cases I would question the second claim, that they do not deserve praise. What is it about children that makes them improper targets for praise? They are capable of

accomplishments, of succeeding where it is difficult and of overcoming obstacles. These are all proper occasions for praise. They are not, however, necessarily occasions for moral praise, and that is what is at issue. Do children ever deserve moral praise? Sometimes they do, but sometimes they do not. Presumably to have morally worthy qualities of will it has to be the case that one has some basic understanding of morally salient facts that children below a certain age simply don't. So it is often said that children below a certain age are completely egocentric, in that they do not know that other people have intentional states, can suffer harm, etc. It is hard to see how someone so cognitively underdeveloped could count as a moral agent, and so hard to see how they could deserve to be praised morally. The question then becomes whether it is legitimate to praise children morally when they are in this state.

The answer, at least for my part, is no. And I can't see any reason to think it is. It is useful here to remember that praising and blaming, on the **Quality of Will Account** are not identified with any public actions. Fundamentally they are the private actions involved in making certain kinds of judgments. So even if we have a tendency to say the same thing to a child when she shares her toys that we say to the adult who makes a charitable donation, that does not mean we are actually praising the child. We are acting as though we praise the child. There are many benefits to doing so. It is a good way to reinforce good behavior on the child's part. And because we are not actually praising her there is nothing unfair happening here.

What about cases where people are rewarded for victories that are matters of luck? Lottery winners are rewarded for winning, and it is not clear that they deserve the money they are given just because they make a lucky guess. But it isn't unfair for them to be rewarded this way. So we have an undeserved benefit that does not constitute an instance of unfairness and thus a counterexample. Or so it appears. In fact lottery winners very well might deserve the money they receive. They were, after

all, promised that if they guessed correctly then they would receive the money. As long as the promise met certain conditions, once it was made it became true that the winner would deserve the money. What conditions are these? Well the promise needs to have been made by a duly authorized agent of the government, or whatever institution is giving the money away. Presumably the promise can't be an instance of deep irrationality or evil. If the promise was made by someone who thought that in making someone wealthy without having to work for it they would destroy the person's character then perhaps the promise has no desert-making force. That is why I commit myself to saying no more than that the lottery winner might deserve the money.

But what about the intuition that the lottery winner doesn't deserve the money? Are we to think of it as simply the expressions of a confusion? I think that the intuition expresses the fact that the lottery winner does not have a basic desert claim to the money, along with a prejudice against anyone getting anymore than they basically deserve. What is it to basically deserve something? To deserve it but not because some person or group decided that you deserve it. Sometimes we do deserve things simply because someone else decided we deserve it. When we are the beneficiary in a will we come to deserve what was willed to us just because the deceased said we did and certain defeating conditions are not present, such as it being the case that the deceased never had a right to give the bequeathed object away. When in the midst of a dispute the disputants submit themselves to arbitration, the arbitrator can make it the case that one of the disputants deserves something by deciding that it is so. Just like the case of the will, the arbitrator can only do this when certain defeating conditions are not present, such as the arbitrator having a conflict of interest. Something similar might go for governments, assuming that their procedures and laws are just. Basic desert, I am saying, cannot arise in the way that the desert in these cases arose. X

cannot basically deserve Y just because someone said so, no matter who that person is. In what sense is basic desert basic? Well in the cases of the arbitrator, the will, and the state, the desert could only be generated when certain defeating conditions were not present. Those defeating conditions all had to do with whether the person generating the new desert themselves deserved to be making that decision. The desert generated by the decision derived from the fact that the person making the decision deserved to be in that position. So basic desert is to be contrasted with derived desert, and the desert that the lottery winner has with respect to his winnings is derived desert. The lottery winner deserves the money only because he was promised that money by a proper authority.

The intuition that he doesn't deserve the money might simply come from a confusion. It might be that people are confusing the lack of basic desert with the lack of desert. But the roots of intuition probably go deeper. There is, on the part of many, significant resistance to the idea that someone ought to be successful in life simply through the decisions of others. While few people would endorse the claim that there is no derived desert, many people think that derived desert is very restricted, or at least that the more a person's life is one they basically deserve, the more admirable or choiceworthy their life is. I think this dislike of pervasive derived desert leads to people not wanting to acknowledge the derived desert, and so treat cases of derived desert as though there is no desert present at all. While I am in agreement that derived desert should not pervade human affairs or even a single person's life, I think we should resist the temptation to deny it altogether.

So while it has some prima facie odd consequences, the principle that it is unfair to harm or benefit someone who does not deserve to be harmed or benefited is plausible. The principle that is hard to accept is **PBBH**. This principle would not be hard to defend if I endorsed an account of praise and blame in which they were a kind

of public sanction or reward. Everyone realizes that it harms people to be denounced in public and that it helps them to be lauded. But praise and blame are not public rewards or sanctions. Sanctions and rewards can express blame and praise, but blame and praise need not be expressed this way. They need not be expressed at all. Given the private nature of blame and praise, and given that we can blame and praise complete strangers and the long dead, how could **PBBH** be true?

4.2 The Case for PBBH

I will start out by admitting a reason you might doubt **PBBH**. It seems to grant perfect strangers incredible power over me. According to **PBBH** if someone on the other side of the world that I have never met and never will meet blames me for something, they have managed to harm me. This despite the difference, despite the fact that I could not possibly find out, and despite the fact that we have no relationship with one another. How could this possibly be? What aspect of my life is negatively impacted by a private mental event in the head of someone on the other side of the world? How could an event which does not have any causal effect on how my life goes and is completely unknown to me harm me, or benefit me?

Let me start off by pointing out that the potential privacy of praise and blame does not seem to be an obstacle to it being good or bad for us. Consider the following example:

> *Insulated Father*: Jack very much wishes that his daughter would love and respect him. He takes his role as a father to be the most important aspect of his life, and he would consider himself and his life a failure if his daughter did not love and respect him. Jack's daughter is extremely polite and kind, and would never, even if she did not respect or love her father, let him know that this was true. Jack's daughter is extremely accomplished at hiding her true emotions, and as a result Jack can never tell whether his daughter really loves and respects him, or whether she merely thinks he deserves to be treated kindly.

It seems to me that Jack's life goes better if it is the case that his daughter actually

loves and respects him, but because of his daughter's tendency to express only kind, polite emotional responses, whether or not she loves him is something he cannot come to know. The privacy of the his daughter's attitudes toward him do not present a barrier neither to her hating him being bad for him nor for her loving him being good for him. If Jack were given the choice between a world where his daughter loved him but he never found that out and a world where his daughter despised him and he never found that out, does anyone doubt that Jack has good selfish reasons to prefer the former? His life goes better in the world where she loves him because it means that a large and significant project of his, that of fatherhood, has not been in vain.

Privacy does not seem like it could be the reason why **PBBH** should be rejected.[139] What about causal isolation? Someone in Siberia who somehow hears of me and something I did could blame or praise me. While I had some causal effect on him, there is no way for his blaming or praising me to have a causal effect on me or any part of my life.[140] So how could it be that he could hurt or help me? To see that causal explanation is not a problem consider the following example:

> *Charles the Classicist Monk*: Charles is a monk in the early middle ages. Charles has a deep and abiding love the classical works of Mediterranean civilization and sees that those works of art, science, literature and philosophy are under threat. So Charles takes on the project of protecting as much of classical tradition as he can. He spends all his time looking for classical works and finding places to hide them. Charles has the ability to be this single minded because he is the abbot of a monastery, and so no important duties of his are being neglected. Charles' goal is to save as much as he can in a safe place that will only be found once the long train of uncouth barbarian invaders stops and the coarse Germanic hordes settle down and come to appreciate all that he has saved. Charles realizes that this day will only occur, if it does at all, long after his death. Charles dies with this

[139] According to some accounts of the good, privacy just has to matter. On hedonistic and other mental state views the fact that Jack cannot find out means that his daughter's private attitude towards cannot be good for Jack, except indirectly. I argue against such views later.

[140] It might seem like I am begging the question against my own view, but I am not. IF **PBBH** is correct than the Siberian blaming me just is harming me. The relationship between the blame and the harm is not causal, it is logical or conceptual.

> treasure of old books and scrolls safe, not knowing how things will turn out.

I think that given Charles' goals, his life was a better life if it is the case that his treasure trove is actually discovered by people who appreciate it, rather than being discovered and burned by illiterate barbarians. The reason why it would be better is because on one outcome his life was lived in vain, and on the other it was not. By adopting the project he did Charles brought his own good into close connection with something that would outlast him. By organizing his life around the goal of making sure that facts in the far future obtained, he made it the case that if they did not come about, the actions in his life that were in service of achieving his goal were pointless. If this is right then for some people events that occur after they die can make a difference to how well their lives went. It is hard to imagine a person being any more causally isolated from an event than by being deceased.

So the two most obvious reasons to deny **PBBH**, the privacy of blame and the potential causal isolation of blame, are not in general barriers to a state of affairs being good or bad for a person. Are they problems more specifically for praise and blame though? The question is pertinent because while causal isolation, for example, is not generally an obstacle for something being good for a person, it is an obstacle for certain states of affairs being good. Food with good flavor, for example, is good for no one if it is causally isolated from everyone. Words of encouragement is good for no one if they are kept private. Why aren't praising and blaming like that? Put in other words, flavorful food seems good because it makes possible the experience of tasting flavorful food, and words of encouragement seem good because they can actually encourage those who hear them. The question is whether praise can only be good because of how it contributes to the experience of being praised, and blame only bad because of how it contributes to the experience of being blamed.

I do not think this is the case. I think it is obviously not the case for praising and blaming done by certain people. If those who we care about or respect blame us that is bad for us, and if they praise us that is good for us, whether or not we know or could ever know. Of course proving that this is so is difficult. One cannot after all bring someone to appreciate the fact that something they are incapable of knowing about is good or bad for them. But I think there is some evidence for the claim, even if it is not dispositive. Suppose you hear from a friend that a shared acquaintance, of whom you think very highly, believes that you live a trite meaningless life. Certainly this will be distressing. Why will it be distressing? Those who question **PBBH** might be tempted to say that this is distressing because of how unpleasant the experience of finding out was. Or they might wish to say that this knowledge now makes it impossible for you to interact with the shared acquaintance in anything like the way to which you have become accustomed. These explanations surely explain some of why we would never want to face the situation of finding out that someone we respected did not respect us, but I do not think they explain everything. While it is true that we would not want to find this out because the finding out is unpleasant, we still need to know why it is unpleasant. What makes not being respected by one you respect so bad? And the prudential concern one has for the smoothness of social engagements does not seem to explain why it is, upon finding out that someone thinks ill of you, that you feel as though something was wrong before finding out. It is galling to think about a meeting in the past in which you now know the person you were meeting with thought poorly of you. It is not always galling because you think they might be right. Nor is it always galling because you think they are wrong. It is sometimes just the fact that they were thinking poorly of you then that you now find so galling. It is hard to see why, if your concern is with future encounters with a person, finding out that they think poorly of you should lead to your being displeased about past encounters with

them.

The best explanation of what is going on in cases like that of the respected acquaintance who does not think much of you, is that you have a desire for the people you respect to think well of you. When you find out that the desire has not been satisfied, you are upset about that fact, a fact which was true in the past and will be true in the future. If I am right, that still does not provide an answer to the question of whether or not being blamed is bad for us in a way which allows for it to be bad for us even though it is private and causally isolated. All I have shown, if I am right about why blame of any kind from those we care about is upsetting, is that we do not desire to be praised and not to be blamed in the way we desire flavorful food. We do not desire praise because the experience of praise is so nice. We just desire that we be praised.[141] But that does not establish that praise is good and blame is bad, just that if praise is good it is good in a way that is immune to the worries of privacy and causal isolation, and so also with blame. But it might be that praise is not good for people and blame not bad for people, because it might be that the desires I claim we have are desires we ought not have. This is just the application of the well known point that just because you desire something does not make it good for you.

I think that not only are these desires not desires it is bad or wrong to have, they are desires that we ought to have. In fact I think that we ought to care whether everyone else, not just friends and loved ones, praises or blames us. I think this fact, that we ought to care about what everyone else thinks of us, morally speaking anyway, is strong evidence for **PBBH**. There are two ways that our having such a duty would be evidence for **PBBH**. The two ways correspond to two different relations that might obtain between obligations and intrinsically good states of affairs. It might be that the

[141] To avoid sounding as though I am endorsing superficiality, let me remind the reader that praise, on the account I am considering is a judgment about someone's quality.

explanation of why it is that we ought to care about what others think of the justifiability of our actions is that their praising us is intrinsically good for us and their blaming us is intrinsically bad for us. If this were the case then **PBBH** would not only be a true moral principle, it would be morally or ethically basic. Alternately it might be that the fact that we ought care about what other people think of the justifiability of our actions is what makes it the case that praise is good for us and blame is bad for us.

There are moral theories according to which one or the other of these two explanations must be false. Some extreme deontologists[142] might be tempted to say that we never have a moral duty because of the intrinsic goodness of some state of affairs that the duty directs us to pursue. Some extreme consequentialists might think that it is impossible for obligations to hold except for when they are obligations to promote intrinsically good state of affairs. Those who subscribe to one or another extreme view can simply pick one kind of explanatory priority and not the other. I subscribe to neither and I can't imagine why anyone would want to lessen the versatility of legitimate ethical thought by endorsing one of the extreme views. As long as everyone is going to admit that one of these two relationships between intrinsic goodness and obligation could obtain, I am pleased to move on to objections to the claim that if we ought to desire to be praised and not to be blamed then we have good evidence that praise is good for us and blame is bad for us.

First to why it is not wrong to desire that others praise us and not blame us. On the face of it such a desire appears shallow. It is received wisdom that people who attach a great deal of importance to what people think of them generally are shallow. So it is worth noting that **PBBH** is not meant as a principle covering every kind of

[142] So extreme that this characterization of deontology might be so narrow as to exclude Kant, given his commitment to the intrinsic goodness of rational nature and the role that goodness seems to play in some of the formulations of the Categorical Imperative.

praise or blame. **PBBH** is meant only to apply to moral praise and moral blame, and there are other kinds. We can praise someone for their artistic work without ever making any judgments as to the moral worth of her character or of her decision to create the artwork. We can praise people athletically for their feats of strength or speed or agility. While it might seem that we do not engage in artistic or athletic blame, that is, I think, only due to linguistic accident. We do something very like blaming in athletic cases, when we make judgments like 'It is Bill Buckner's fault that the 1986 Red Sox lost the World Series.' This is a judgment that assigns fault for an event that is significant, at least in the context of discussions of sports, or in any context in New England. And it is not just the assignment of causal responsibility, because there are many factors that were equally causally relevant to the Red Sox losing the World Series that year. So, for example, we do not think it was the general manager's fault that the Red Sox lost because he signed Bill Buckner to a contract, even though Buckner's mistake would not have been possible without that event. We assign blame to Buckner because as the first baseman it was his task to catch grounders headed to first base, and had he done it the Red Sox would have won the World Series.

I am not claiming that we should care about artistic, athletic or any other form of praise and blame besides the moral form. It would be shallow to care what perfect strangers thought of your shoes or your hair, but that does not mean it is shallow to care about what people think of the justifiability of your actions.[143] And we are all already committed, I think, to the view that some people's judgments about the justifiability of our actions are things that we ought to care about. We ought to care about what our friends think of the justifiability of our actions, because not to do so

[143] I think the level of shallowness corresponds to the objective significance of the goods in question. Certainly there is some good involved in dressing well, perhaps even a good of an artistic or aesthetic kind, but it is certainly not among the more significant aesthetic goods, and there is no reason to give much thought to what random strangers think about it.

would be to treat the friend in a way that she does not deserve to be treated, given that she is a friend. Most of us also think that it would be better for us if people we consider moral exemplars thought that our actions were justified. These desires do not express any shallowness on our part. Those we care about and respect are those whose moral opinions we care about. Sometimes, as with respecting someone as a moral exemplar and perhaps with friendship, caring about their opinions in this way might be a constitutive feature of the relationship, while in others, such as that of family member it is simply an excellence of the relationship which is not essential to the relationship. If there is a problem with desiring the praise of others and desiring not to be blamed, it is a problem with desiring this with regards to certain kinds of people, not a problem generally. The two most problematic cases seem, to me anyway, to be praise and blame from strangers and praise and blame from the morally perverse, which I will examine soon.

An objection related to praise and blame from strangers is that if there were a duty to care what everyone else thought of the justifiability of our actions, and we obeyed that duty, then we would make our own good far too dependent on the whims of others. Imagine if some action of mine were held up for public scrutiny, such that everyone in the world was aware of it and able to form judgments about it. Suppose that in this case all of them did form judgments and they all formed negative judgments. All of the sudden according to **PBBH** 6 billion bad things just happened to me. Shouldn't 6 billion bad things happening to you ruin your life? But isn't it absurd to think that your life could be ruined because of 6 billion private mental events? Imagine that you did not know that your action was publicized, and were just going about your day, when, all of the sudden, your life went from good to bad because 6 billion people just decided that you were a mean person. That sounds perfectly absurd.

I admit that if my view implied that X's life could be ruined because a bunch of strangers thought ill of him, then my view would be false. But there are ways to avoid that implication. One way is to simply grab the other horn of the dilemma and claim that the harm involved in being blamed is infinitesimally small. While this would provide a way out it would undermine the larger view, which is that principles of fairness can and do apply to praising and blaming. If the harms involved in blame are infinitesimally small, then it is not plausible that it is unfair to inflict that harm on someone. Fairness is not at issue for completely trivial benefits and harms, which is what praise and blame would amount to if blame were so inconsequential that no amount of blame could ruin someone's life.

I prefer to pursue other another option, which is that there are non-trivial harms which are such that they cannot ruin someone's life, no matter how often they happen. There are many examples of such harms. Most of the examples share with blame and praise the property of being harms that a person need not be aware of to count as a harm. So I take it that I am harmed to a small but not trivial extent when I miss out on some opportunity for fun. Things go worse for me if I pass up a chance at an enjoyable party, or a relaxing day in the sunshine. These harms are not trivial, because if someone else were to force me to suffer those harms when I did not deserve to suffer them, they would obviously be treating me unfairly. If I am denied access to a party because of my race or gender, then even if there are other fun things to do, I have lost an opportunity and have been treated unfairly. But presumably no amount of missing out on opportunities for small goods could ruin my life. After all I miss out on those opportunities constantly. There are, each day, many different ways I could spend my time, many of which would be good for me. But the aggregate of these small losses cannot, no matter the number constitute a harm that would ruin my life. They cannot even make it the case, all by themselves, that my life goes significantly

worse. I think that blame from strangers is such a harm. It is non-trivial so it is something we should not inflict on someone when they do not deserve it, but it cannot lead to a person's life going significantly worse.[144]

But why should we care about praise and blame from people we have never met in the first place? It matters to us what friends, loved ones, colleagues, and personal moral heroes think of our actions because those people matter to us. But strangers do not matter to us, at least not in a similar way. It is not as though we bear the same relationship to strangers that we do to friends but to a lesser degree. In arguing against Scanlon I appealed to the fact that a relationship seems to require interaction, and as that is absent in the case of strangers there is no reason to think that we stand in any relationship to them which would explain why it is that we ought to care what they think about the justifiability of our actions. The reason why we ought to care about whether strangers praise or blame us is not a relationship they stand in to us, but features of theirs. We ought to care what they think about the justifiability of our actions because we ought to care about their moral opinions more generally, and we ought to care about their moral opinions because that is part of what it is to treat them as moral equals.

It is important to distinguish between two different reasons why we ought to care about the moral opinions of other people. One is that other people are a potential source of moral information in the same way they are potential sources of non-moral knowledge. If you want to know how to do something, take care of a horse for example, you should listen to the opinions of those who spend a lot of time dealing with horses. Similarly if you want to know what is right and what things are good, it is a good idea to listen to those who frequently have to deal with such questions.

[144] At least not in general. If someone decides that being well thought of is her goal in life then perhaps blame from strangers could make her life go significantly worse. But this is because the harm has been amplified by decisions she has made about her life projects, not because of the harm inherent in blame.

Because almost every adult human being has to deal with those questions, we have prima facie reason to care about what every adult human being thinks about the justifiability of our actions if we have reason to want to increase our moral knowledge.

There is a second reason, one that yields a pro tanto reason to care about what other people think of the justifiability of our actions rather than a prima facie reason. Every person might, for all we know prior to meeting them, be a source of moral knowledge, but every person is a creature that deserves a certain kind of regard and respect. Adult persons should not be treated as something they are not, and to treat someone as though their moral opinions were of no relevance would be to treat them as something they are not. Small children are creatures whose moral opinions we need not take seriously. The reason why is that small children are cognitively and emotionally limited in certain ways that make it plausible that they are not our equals when it comes to moral reasoning. That is why one does not treat a child unfairly when disregarding their opinion that a fair game would always end with them winning. But if we treat adult persons this way we are acting as though they are similarly cognitively and emotionally limited, and to treat adult human beings as though they had the cognitive and emotional capabilities of child is insulting.

Of course there are some people whom we are permitted to not treat as moral equals, other than children.[145] Sometimes people, through their actions or their avowed intentions and values, or both, show themselves to be so deeply morally mistaken that it is not longer reasonable to treat them as though they are your moral equals. Or, even if they are still moral equals, their moral thinking is so perverted that it is no longer worth caring about. Think here of the Neo-Nazi or the serial pederast.

[145] To be clear, I think the case of moral perversity is a case in which the pro tanto duty to care about the morally perverse person's moral opinions is outweighed by other factors. The fact that there is a pro tanto duty to care is not enough to save **PBBH** because it might be that the reason why the pro tanto duty is outweighed is because the pro tanto goodness of the praise of the morally perverse person is outweighed by some set of pro tanto harms involved in their praising you. In such a case the praise of the morally perverse would be a harm.

Is there any reason, having to do with either looking for new evidence about morality or with showing people proper respect, to think that the opinions of such people about the justifiability of our claims is relevant? There are two ways to deal with this problem.

One has to do with the nature of incompatibilism. Incompatibilists are all committed to the conditional 'If determinism obtains then no one is morally responsible for what they do.' Assuming, for the sake of the argument, the **Quality of Will Account**, what the consequent amounts to is 'none of the moral judgments involved in blaming and praising are appropriate.' If all that incompatibilists are committed to is the claim that if determinism is true then none of the judgments involved in blaming and praising someone are appropriate, then it doesn't look like the praising and blaming of the morally perverse needs to be treated as a benefit and harm. After all **PBBH** is needed to get from the **Quality of Will Account** to fairness based incompatibilism, by showing that for praise and blame to be legitimate the judgments that constitute them would have to be fair. But when the morally perverse praise and blame people, the question of whether the praise and blame is appropriate can be answered without appeal to questions of fairness. What is uncontroversial is that it is not appropriate to make a judgment on the basis of poor or no evidence, and this is presumably what the morally perverse person does.[146] So why shouldn't the incompatibilist say that the deeply false beliefs that the judgments of the morally perverse either wholly or in part express make those judgments inappropriate, and that

[146] It is not plausible that anyone is always wrong morally. After all even Neo-Nazi's need to get along with each other, and so they are going to respect rules of fair play in some contexts. What is going to be true of the deeply morally perverse is that the moral principles they accept are going to be corrupted in such a way that even when they get the right answers, the judgments they support will not be well supported. Take the Nazi who thinks that it is unfair to, for example, take a promotion within the organization that one has not earned. The Nazi might even think that to do so would be unfair, but it is worth asking what conception of fairness this could be, which could allow the murder of millions of people. Would a principle of fairness that looked something like 'Give a pure blooded white man his due' count as adequate support for a judgment of fairness? I don't think so. This corruption of even right answers is part of what sets the morally perverse person off from the simply morally confused.

for other people whose moral values are not deeply corrupted, considerations of fairness imply that if determinism is true then none of their judgments would be appropriate? This would still count as an argument for the claim that if determinism is true then no praise or blame is appropriate. What this approach gives up on is the principle that determinism is what makes all praise and blame appropriate. The position I am outlining is one on which **PBBH** is only meant to apply to praise and blame from the non-morally perverse and so that the inappropriateness of praise and blame would, in a deterministic world, sometimes be explained by the fact that determinism obtains and sometimes by the deep moral perversity of the individual making the judgment.

There is another way of dealing with the problem of the morally perverse that does not require a the reinterpretation of incompatibilism I just mentioned, and it is to treat the morally perverse as though they have praise and blame reversed. After all, to find out that the Neo-Nazis approve of you would be disturbing, but their hatred for you and what you do is a kind of honor. So why not think that in the case of the morally perverse praising is a harm and blaming is a good? Certainly they don't mean to be praising you when they say 'You are a race-traitor' but most of them also don't mean to express false moral views, so I don't think this is much of an obstacle. As with the previous suggestion this response to the problem of moral perversity requires a revision to **PBBH**. In the first solution the scope of **PBBH** was restricted to cases of praise and blame by the non-morally perverse, with praise and blame by the morally perverse being ruled out for reasons not having to do with determinism and fairness. This solution also revises **PBBH**, but not in a way that fundamentally changes the case for incompatibilism. So the new version is:

> **PBBH**: If X praises Y, X always either benefits Y (the standard case) or harms Y (the moral perversity case) and if X blames Y, X always either harms Y (the standard case) or benefits Y (the moral perversity case).

If correct this revised version of **PBBH** gets around the problem of moral perversity. The idea is that in the case of moral perversity it is not that we do not or ought not care what the person thinks, it is just that our caring about what the person thinks gets expressed in a different way. We ought not to want to be praised by the morally perverse and we ought to be proud that the morally perverse blame us.

The question is whether **PBBH** is true. In answering that question it is important to keep in mind the account of praise and blame under consideration. **PBBH** is about the making of judgments, not the expressions of those judgments. To see the importance of the contrast, consider the famous case of the Nazis marching in Skokie, Il. In the course of the march a lot of expressions of disgust, hatred and blame were made, and it would be absurd to think that those expressions were good for the people at whom they were directed. Many residents of Skokie were Holocaust survivors and the public expression of the ideology that was used to justify genocide was traumatic. It is here that it is important to distinguish between the blame, which is constituted by a judgment made by the Nazis that Jews are guilty of all kinds of absurd conspiracies, and the expression of that blaming judgment. The expression of the blame is clearly bad for residents of Skokie. That does not imply that it is bad for them to be blamed, though. Does it do them any good to be blamed by the Nazi's though? It certainly wouldn't cause them pleasure to think about the fact that there are virulent anti-Semites, though perhaps this might be explained simply by the painful thoughts with which anti-Semitism is associated.

Still, it does seem like the correct response that Holocaust survivors ought to have to the fact that some Neo-Nazi's make negative moral judgments about them is

one of indifference.[147] I have been using Neo-Nazi's as clear cases of moral perversity, and treating all forms of moral perversity similarly, but in reality the situation is more subtle than that. To see some of the subtlety imagine two different Neo-Nazi's. One is irrational in his racist beliefs and behavior. He is not able to support his positions with arguments, and when pressed for evidence or reasons he responds with either ad hominem attacks or incoherent conspiracy theories. The other Neo-Nazi has a coherent moral theory, which he explicitly employs to justify his actions. At certain key points his moral theory is mistaken, and he is culpable for his ignorance in part because it seems clear that he has chosen the moral theory he has in part because he simply enjoys the suffering of those weaker than him. The first Neo-Nazi is one whose opinion we ought not care about. I think the explanation is that his moral thinking doesn't meet basic standards of rationality and more specifically that his blaming and praising do not qualify as rational responses to the actions of those he is blaming and praising.

Does the case of the irrational blamer mean that I am forced back to the first strategy, in which the scope of **PBBH** is restricted and reasons other than determinism are required to show that if determinism obtains then all praise and blame are inappropriate? I do not think so, because I think that for a moral judgment to count as blame or praise, it has to make sense as a response to the action it is about. To make sense as a response to an action there has to be some rational connection between the action and the judgment that constitutes the blaming or praising, such that other rational individuals can comprehend how someone would make that judgment about the action. So, for example, no rational person would conclude, from some other person leaving a small tip, that the bad tipper was a heartless monster to the waiter.

[147] I do not mean to say that Holocaust survivors, or their descendants, or any decent person really, ought to be indifferent to the fact that there are Neo-Nazis. This should horrify people. What seems not to deserve a second thought is what the Neo-Nazi's think, except insofar as knowing what they think could help bring it about that there were no more of them.

Perhaps the bad tip was simply being used as an occasion to publicize one's long standing feelings about a person, in which case the blame is actually about some other actions other than the tipping. But if the judgment is not an expression of some potentially well grounded judgment than it looks like it is just the person being mean, not actually evaluating the justifiability of the tipper's action.

Another constraint on blaming and praising that you get out of this demand for minimal rationality is a constraint on how much the judge's description of the non-moral properties of the action can diverge from the description of the non-moral properties of the action on which the agent acted. To see why, consider the action of Jon helping an old lady cross the street. Suppose Jon thought it worth doing because she is an old lady and they often have trouble with their mobility, and that she deserves a helping hand after all that she has likely been through. Tyler is watching from afar and for some reason concludes that Jon has kidnapped an old woman. If Tyler were to think of himself 'Shame on you Jon for kidnapping!' would this be a response to what Jon actually did? No, because Jon did not kidnap anyone. You cannot blame someone for what they did not do. The less divergence there is between the descriptions of the actions, or between the descriptions of the action and what actually happened, the harder it gets to tell whether or not the judgment counts as blame or not. In the case of the irrational Neo-Nazi it seems clear though. Suppose part of the explanation of anti-Semitism is the confusion of the insular nature of many Jewish communities due to rules about diet and marriage for sign of a cabal bent on controlling the world. When confusions like this are present the moral judgments that arise from them do not count as actual blame. No one is blaming Jews for anything when they say that Jews ought not keep such a tight control over the international banking system. Nazis can attack Jews, insulting Jews, but try as they might the Anti-Semite cannot actually blame Jews for that, because it is a fiction.

So when the **Quality of Will Account** is presented properly some of the clearest cases of where we ought to be indifferent to praise and blame turn out not to be cases of praise and blame at all. Taken together with some form of **PBBH**, the **Quality of Will Account** does not rule out fairness based incompatibilism. If you supplement **PBBH** and the **Quality of Will Account** with, for example, some principle of fairness which requires that people be able to avoid harms or fail to receive benefits for them to be legitimately given, then you have an argument for incompatibilism, at least on the most widely accepted accounts of what it is to be able to avoid something. The argument hinges on the truth of **PBBH**, a principle I have tried to make look plausible and intuitive. If I could do that much I would be happy, but any case for **PBBH** that only did this would not give us good reason to think it is true. **PBBH** is a principle about what is good, and what is not, for people. So **PBBH** ought to be at least consistent with some plausible accounts of the good, and the more the better.

4.3 PBBH and Accounts of the Good

The first theory of the good that I will look at is Hedonism. Earlier, I raised as a potential problem for **PBBH** is that it says that potentially purely private events in the minds of other people can be good or bad for us. Praise and blame, being private actions, are things that the objects of the praise and blame are not aware of, unless the praiser/blamer decides to express the praise and blame. So, if it were the case that all harms and benefits had to be experienced by the subject of those harms and benefits, then **PBBH** could not be true. What reason might there be to think that all harms and benefits have to be experienced? One historically common reason to think it is that harms and benefits must be experienced is that only conscious states, or objects which causally produce conscious states, can constitute harms or benefits. Hedonism is a theory of the good which is committed to this claim. If pleasure is the only thing that

makes a person's life better and pain is the only thing that makes a person's life go worse, as hedonism says, then praise and blame just cannot count as goods and bads.[148] But is hedonism an adequate account of what is good for people?

No, it is not.[149] Hedonism is subject to a wide range of counterexamples to which no effective response has been given. What is more these counterexamples are all either counterexamples to the claim that harms and benefits must be experienced, or can be made to stand as counterexamples with minimal revision. The most recent of the classic counterexamples comes from Robert Nozick and his experience machine example.[150] In this example we are asked whether it would be in our best interest to hook ourselves up to a machine that could flawlessly simulate all the experiences we want in life. None of these experiences would be veridical. Rather than having a deeply fulfilling family life, or a successful career, or deep knowledge of important subjects one has experiences that perfectly replicate what it would be like to have all those things. One is actually lying in a vat of goop that somehow facilitates the production of non-veridical experiences. Nozick thinks that it is not in our best interest to hook up to the machine. I agree with him. The life spent in the vat of goop is almost entirely without value. The prospect of living such a life is horrific.[151]

What is so bad about it? Well for one thing one is massively deceived. That by itself is enough to make the life bad. But more importantly the life in the vat of goop is one devoid of accomplishment, success, friendship, love and the like. It is

[148] I have switched from talking about benefits and harms to talking about goods and bads on the assumption that if X is a benefit then X will be recommended by any adequate account of the Good.
[149] Despite the fact that disagree with him on almost every point I am following Fred Feldman's (2004) organization and understanding of the objections to which hedonism is liable.
[150] Nozick (1974) pg. 42-45
[151] Though it ought to go without saying let me assure the reader that I am not imagining being hooked up to the machine but being aware of that fact as I am being fed wonderful experiences. Fred Feldman (2004) seems to think that philosophers who do not share his judgment that the vat life might be good for the person having it might be making this very elementary mistake in thinking about the example. I am not guilty of doing so and I both seriously doubt any other person trained in philosophy is and am quite certain that we do not need to consider such mistakes to explain the intuition that the vat life is bad for the person having it.

bereft of all of the things that would make one's life worth living. Far from guaranteeing one a good life by guaranteeing good experiences, hooking up to the experience machine guarantees that one will fail to have any of the really significant goods.[152]

Another classic objection to hedonism is that it treats the pleasures that a person experiences in the course of performing base and depraved actions as good for the person performing those actions. On the face of it this seems deeply wrong. After all it is bad for people to live lives that are full of morally, prudentially and aesthetically inferior actions. It is bad for the drug addict to give in to their addiction, it is bad for people to maintain deceptive and fraudulent facades for the purpose of the cheating others. Feldman, in discussing this objection, gives us the example of the revolting action of a human being having sex with a pig. The point is that it is not only morally wrong to take pleasure in certain events or states of affairs, it is bad for the person taking that pleasure to take pleasure in it.

The final objection I will mention, although there are others, is what Feldman calls the 'shape of life' objection. This objection, offered by Velleman and Slote, consists of the comparison of two lives, equal in amounts of pleasure and pain experienced. Given that information we already know that hedonism, of any straightforward variety, is going to have to treat the lives as equally good for the people who lead them. But, Velleman and Slote ask, what if those pleasures and pains are distributed differently within the two lives? What if one of the two people starts life with a significant amount of pain and not much pleasure, but through hard work makes his own life more and more pleasant and less and less painful, while the other

[152] I say significant goods because clearly the person in the vat has some good things in her life. To use a personal example I love the feeling of sunshine on my back on a very still day. It is not particularly important to me that the day be clear or even that it be sunshine. This is just typically how I get that feeling. If that feeling could be recreated it would be just as good for me, to feel the artificial sunlight as the real sunlight. This is a good and it is a good available to me in the vat. But I would not trade one day with my daughter for a lifetime of sunny days, and so I think the good is not very significant.

starts out life with a great deal of pleasure and through inattentiveness or laziness feels less and less pleasure and more and more pain as life goes on? Isn't the first life better than the second life? Wouldn't anyone interested in living a good life rather live the first rather than the second life? Wouldn't you wish the first life rather than the second for your child?[153] As far as hedonism is concerned these lives are of equal value, because they contain the same amounts of pleasure and pain. But here, as with issues of social justice, distribution matters.

Do hedonist's have any responses to these objections? Yes. In fact responses are relatively easy to come by if the hedonist is willing to tolerate ad hoc additions to this theory which strip it of all the explanatory power and simplicity it started with. Feldman for instance has suggested maintaining a commitment to the claim that pleasure and pain are the only things that can make a person's life go better, but simply saying that only certain kinds of pleasures and pains can do that. So if a hedonist were willing to say that only pleasures taken in worthy objects and pleasures taken in things that really happen can make a person's life better.[154] These additions to the theory easily take care of the objections having to do with unworthy pleasures and with the objection concerning Nozick's experience machine. How could they not, they were added to the theory precisely to get around them.[155]

---

[153] I present the objection simply, but it is not at all clear to me that the choice is clear given these simple descriptions. Assuming that it is reasonable to test theories of well being by asking one's self 'What would I want for my child?' I find that I am hard pressed to choose the life in which my child feels a great deal of pain early in life. That is, after all, when she will be least able to handle the pain. But this concern is not relevant to the debate. I am not contesting the claim that distribution of pleasures and pains matter. In fact I affirm that claim in expressing my concern with giving my daughter the life with the upward trajectory. All I am doing is objecting to the particular distribution at issue. If we were comparing two lives that contained blissful childhoods, and only start to diverge at adulthood with one life suddenly filled with a great deal of pain that is slowly but surely eliminated, while another life begins an inexorable downward path, I would choose the first over the second for my daughter.

[154] Feldman (2204) endorses attitudinal hedonism in which the pleasures which are the source of well-being are not sensations but attitudes directed at states of affairs. So when the relevant pleasure in the case of the drug addict is the attitude expressed by 'I take pleasure in the fact that I am shooting up now'.

[155] Feldman (2004) simply bites the bullet when it comes to the Shape of Life example, again after suggesting that the intuition that one life is better than another might be the result of some very

But ad hoc additions like this rarely help theories in the long run. In this case Feldman's revisions to classical hedonism[156]strip hedonism about the good of any significant explanatory power. After all when we want to explain why a person's life goes well we have to mention not just the pleasure and pain she experiences, but the objective worthiness of the objects of her pleasure and the veridicality of the experiences in which she took pleasure. So part of what makes pleasures good for a person are facts that are not themselves facts about pleasure (or pain), on Feldman's view. Is there reason, once we admit that part of what explains why a person's life is good are facts like objective worthiness of objects of enjoyment or pleasure, to reject the claim that a person's life could be made better by events of which she has no knowledge? If mental states are made good for a person in part because of objective facts which do not depend at all on those mental states, why think that only mental states can be good or bad for a person? By undermining the explanatory status of hedonism Feldman makes it hard to see why one would accept hedonism, rather than some more pluralist view, in the first place.

Do other popular accounts of the good imply that **PBBH** is false, as hedonism does? I do not think so. One alternative to hedonism which appeals to many of the same intuitions as hedonism, the informed desire account of the good, does not. On an informed desire account what is good for X is X getting what X would desire if X were fully informed. Informed desire accounts face none of the problems for hedonism that I mentioned above, but maintain the appeal to those who think that what counts as good for people varies according to their tastes, desires and interests. But because informed desire accounts do not claim that something's being a benefit or a

---

elementary mistakes in reasoning on the part of non-hedonists. My own opinion is that if hedonism has to bite the bullet with regards to the Shape of Life objection then hedonism is just false.

[156] I am not counting his endorsement of attitudinal hedonism rather than sensory hedonism as an ad hoc revision of classical hedonism. For one thing I am not sure whether sensory hedonism is the classical form of hedonism and for another there are good reasons, other than simply finding the quickest way out of objections, to prefer treating pleasure as an attitude rather than a sensation.

harm is a matter of the subject being in a certain mental state, it does not imply, straight off, that praising and blaming does not count as benefiting and harming. If a person would, if she were fully informed, desire not to be blamed and desire to be praised, then being blamed would count as a harm, for her, and being praised would count as a benefit.

What has to be true, assuming the informed desire account of the good, for **PBBH** to be true, however, is that everyone would desire to be praised and not to be blamed, if they were fully informed. Either this could mean that there is a natural desire to be praised and not to be blamed, or it could mean that the knowledge that praising is desirable and blaming is not desirable is knowledge you acquire just by being fully informed. The first is a contentious empirical claim, while the second seems as though it might go against the spirit of informed desire accounts. After all if the informed desire theorist can simply build into the information people must have for their desiring Y to constitute Y being good for them, then the informed desire account seems to lose its connection to what people actually desire and the role that intuitively plays in making something good for them. An informed desire account in which objective desirability could be part of the information could imply that the very same things are good for everyone.

So while informed desire accounts of the good are not incompatible with **PBBH**, they would lose the special plausibility they are thought to have over other more objective accounts of the good if they are so revised as to imply that praising and blaming count as benefiting and harming. So lets move on to some of those more objective accounts of the good. Objective list accounts fit the bill, that much is obvious from the name. Could praise count as a benefit according to Objective list accounts? Yes, of course, because just about anything can count as a benefit according to some objective list account or other. It all depends on what is on the list

of objects that are objectively good for people. What makes an objective list account and objective list account is the claim that there is no single property that unifies the items on the list, other than that of being objectively good. So there is no principle which can tell us whether or not something is going to be on the list.

So could I just stipulate that praise and blame are among the items on the list, that they are intrinsically good and bad for people, respectively, and deny that there is any deeper explanation of why? That wouldn't be plausible all by itself, but I could adduce in favor of this stipulation the examples I mentioned earlier, examples where it seemed like praising or blaming someone was unfair. Along with the claim that the best explanation of that unfairness was that praise and blame were benefits and harms, that would give some reason to think that they should either be on the objective list, or be derivable from some item on the list. But this argumentative strategy would leave the entire burden for supporting the claim that praising and blaming could be unfair on the examples above, and this is precisely what I wanted to avoid.

The problem with hedonism and informed desire accounts was that they did not provide objective enough accounts of the good. Hedonism does not provide one at all, and while the informed desire account is compatible with objective goods, the case for that account would be undermined by admitting them. Objective list accounts, on the other hand, provide no principled way of determining what things are objectively good, it just commits one to there being objective goods. What is needed to provide a theoretical defense of **PBBH** is an account of the good that is objective, so that praise counts as a benefit for everyone and blame counts as a harm for everyone, but that provides a principled way to distinguish between goods that are objective and those that are not. Only with such a view would it be possible to give a substantial theoretical defense of **PBBH**. Does any such account of the good exist?

Yes. One example of the kind of account I am talking about is the

eudaemonist account of the good. It may be optimistic of me to talk about *the* eudaemonistic account of the good, since so many different formulations of that account have been presented. In speaking of the eudaemonist account of the good I do not mean to be saying something about all philosophers who are said to have eudaemonist theories. I do not intend to make any claims about the entirety of ancient ethical thought, and pretty much every ancient philosopher who wrote on ethics seems to takes the eudaimon life to be the good life. I mean to present an account what is good that some, but not necessarily all, ancient, medieval and contemporary ethicists accepted. A brief version of the eudaemonist account of the good is that X is good for Y iff Y is related in the proper way to the ends, functions, or activities that X has in virtue of the kind of thing it is. As stated the account is deeply vague. That is just so that it captures the full diversity of views which get called eudaemonist.

To eliminate the vagueness we would need to know what kinds matter in the definition. I am, for example, a member of many, possibly an infinite number of, kinds. I am a human being, a man, an American, a socialist, a weak-kneed agnostic, a student, a AAA member, a St. Louis Cardinals fan, and a Luddite. We need to know which of those kinds matter for figuring out what counts as good for me. By far the most common answer is that it is the biological kinds that matter, specifically my species. But other things might matter. Suppose that God exists, and that I am a creation of his. This seems like exactly like one of the kind memberships that would matter. Many people seem to think that being an American is among the relevant kinds. I think that is almost certainly wrong, but I need not argue for it here. All that matters is that it is clear that any adequate eudaemonist account is going to have to pick a kind or some set of kinds that are similar in a relevant way. Otherwise too many things are going to come out as objectively good and bad for people, and then

the recourse to eudaimonism will have done nothing in the way of justifying **PBBH**.[157]

I am not going to present a fully worked out eudaemonist theory of the good here. All I want to do is show how it would be possible to derive **PBBH** from an account of that kind. If I can show that on one plausibly theory of the good this claim comes out as true, I will have done what I wanted, show how the intuitions I appealed to earlier about Aunt Harriet and others, could fit in an acceptable theory of the good. I am rather in luck, because one of the oldest versions of eudaimonism, Aristotle's version, seems like it has just the implication I am looking for. Aristotle certainly thinks that events completely outside an agent's knowledge can count as harming or benefiting her. He goes so far as to say that events that happen after you die count as helping or hurting you.[158] He also explicitly mentions honor as something that is good for a person.[159]

How does one get from eudaimonism to treating things like praise and honor as goods? By way of answering let me start with the near truism that man is a social animal. To be well functioning a human being, like members of many other species, has to function well in a group. What it means for a human being to function well in a group depends on other features that human beings have qua human beings. In addition to the truism that human beings are social, there is also the Aristotelian truism that human beings are rational. One feature of human rationality is a concern for justification, and that this concern, along with the social nature of human beings, explains, on a eudaemonist account of the good, why **PBBH** is true.

---

[157] The definition needs to be made precise in other ways too of course. Most importantly is how X has to be related to the ends, activities and functions to count as being properly related.

[158] This comment appears in the context of a discussion aimed at showing that the claim that eudaimonia cannot be gained or lost after someone has died, which is a consequence of Aristotle's account of eudaimonia, can be made compatible with what Aristotle seems to have taken to be an obvious truth, that the fate of one's ancestors affects how good one's life was. Aristotle essentially says that some things can be goods even though acquiring them cannot lead someone who was not happy during life becoming so afterwards.

[159] See Nicomachean Ethics Book IV.

It is only with this last step that things start to get controversial. Why does a concern with justification lead to praise and blame counting as benefiting and harming? If we are concerned with justification and we live in community with others, then we ought to be concerned with what others in that community think about what we are doing. To see why this is true consider what would have to be the case if we, while living in a community with other rational beings, did not have a concern with what they thought of us and our actions. In fact this is best thought about by thinking about examples where people have done so. Imagine the theorist who proposes a grand explanation that none of his colleagues can make sense of. Imagine the political activist who is so committed that he doesn't care what evidence people adduce or what arguments they offer against his position. These people either fail to be rational, fail to be properly respectful of their fellow citizens and colleagues, or both. The theorist fails to recognize that his colleagues are in as good an epistemic situation as he is, and the activist fails to realize that being in a political community requires treating fellow citizens as people worth listening to.

If we would care about what other people think of the justifiability of our actions and the reasons why we performed them if we were properly functioning, then that just makes their thinking well of our actions and our reasons a good or benefit for us, on a eudaemonist account of the good of the kind I presented. Presumably others thinking poorly of us would count as a harm because of this. But on the Quality of Will account praising and blaming just are judgments about the moral qualities of actions and the reasons for which agents acted. So on a eudaemonist account of the good of this kind, praise and blame would count as benefiting and harming. Now are there other kinds of eudaemonist accounts? I admitted as much earlier. Would any of them fail to have this implication about praise and blame? Some might, but I see no reason to assume that any would. Those uncomfortable with grounding moral and

evaluative facts in purely natural facts like species membership could find some less natural set of facts to ground their account in, but it would be implausible to deny that to count as well-functioning a human being has to function well socially and meet certain minimum standards of rationality.

Now I have only presented a sketch of a eudaemonist account of the good, because that is all I needed. All I wanted to show was that there was some account of the good which would support the claim that praising and blaming count as benefiting and harming. The details will certainly matter to someone who wants to develop a completely worked out theory of praise and blame, but that is not my goal. I want to show that the Quality of Will account is not committed to compatibilism. If there is an account of the good on which praising and blaming count as benefiting and harming, then the Quality of Will account is not committed to compatibilism. Of course if you were to conjoin the Quality of Will account with a hedonistic account of the good then it would be implausible that praising and blaming are subject to standards of fairness and incompatibilism of the kind I have defended is ruled out. But there is no reason to think that moral theories have no implications about what it takes to be morally responsible, and so there is no reason to be surprised that different moral theories have different implications about what it takes to be morally responsible.

5. Moral Responsibility and Incompatibilism

Now to sum up. Of the three Strawsonian challenges to incompatibilism, one, the Relationship Account, clearly fails. It mistakes certain regular goals and contexts of blaming and praising for the internal goal and meaning of blame. While we often seek, by blaming, to encourage others to change their behavior, we can and do blame people in situations where it is obvious that blaming them will not have this effect. While blaming often has important consequences for our personal relationships, it need not. Incompatibilists need not be worried by the Relationship account of blame

and praise.

Incompatibilists also need not worry about the Emotion Account and Quality of Will Account, not because they are untrue, but because to the extent that these accounts are plausible they are also neutral between compatibilism and incompatibilism. In fact the neutrality of these accounts, when they are properly understood, points us towards a satisfactory account of moral responsibility that can be appealed to by all in the debate about moral responsibility. As we saw there are cognitivist or quasi-cognitivist versions of the Emotion Account. In fact these are the only versions of the account which are plausible. So the Emotion Account when properly understood includes a commitment to blaming and praising having cognitive, or quasi-cognitive content. This dovetails nicely with the Quality of Will Account, according to which blaming and praising is constituted by certain kinds of judgments. And as we saw, incompatibilist challenges enter the picture on both accounts of moral responsibility in the same way, by raising concerns of fairness about the judgments being made.

So what does an adequate Strawsonian account of moral responsibility that is neutral between compatibilism and incompatibilism look like? On this neutral account to praise and blame someone is to have certain emotional reactions to their action. Those emotional reactions are constituted in part or whole by certain judgments about the agent's quality of will in acting as they did. So to legitimately hold someone responsible it must be the case that her actions expressed a quality of will such that it is permissible, i.e. fair, for a person to bear judgment constituted reactive attitudes towards her. Such an account is perfectly neutral between compatibilism and incompatibilism.

Chapter 3: The Moral Luck Argument

In this paper I will present a new argument for the incompatibilism of moral responsibility and determinism. The Moral Luck Argument is new in the sense that no one has ever presented an explicit argument for it, but I take the reasons expressed by the argument to be quite old. They are, I think the standard incompatibilist reasons and intuitions which lie behind other well known incompatibilist arguments. What makes the Moral Luck Argument interesting is that it gives expression to those reasons and intuitions without ever invoking the concepts of 'ability' and 'control' that have caused so much controversy in the literature on moral responsibility and free will. Consider some standard compatibilist goals:

> (A) provide an account of control according to which agents have control over their actions even if determinism obtains.
> (B) provide an account of ability according to which agents have the ability to do otherwise even if determinism obtains.
> (C) show that agents do not require control over their actions to count as free and responsible
> (D) show that agents do not require the ability to do otherwise to count as free and responsible.

What makes the Moral Luck Argument interesting is that it would remain completely untouched even if compatibilists had achieved one or all of A-D. As such the Moral Luck Argument represents a significant dialectical advance over previous incompatibilist arguments.

Luck is becoming a popular topic of conversation in the free will and moral responsibility literature.[160] While luck arguments and objections are typically directed at libertarians, there are some luck arguments against compatibilism. The most prominent and significant example of a luck argument against compatibilism that has

[160] See Mele (2006), Greco (1995), Lackey (2008), Latus (2000), Levy (2008) (2009a) (2009b)

appeared in the literature belongs to Neil Levy. I will start by looking at his luck argument and showing that it both does not present a compelling argument against the compatibilist, and that even if it did his argument would not constitute a significant advance over existing incompatibilist arguments. The problem with Levy's argument is that it relies on an account of what luck is that is both false and strips Levy's argument of any dialectical force against the compatibilist. I will consider two arguments that do not rely on his account of luck, because they do not rely on any particular account of luck at all, rejecting one and endorsing the other.

By way of preparation let me give a very brief presentation of the Moral Luck Argument:

> 1. If it is legitimate to make a moral judgment about a person on the basis of a property they bear which holds of them by luck, then moral luck exists.
> 2. Moral Luck does not exist.
> 3. X is morally responsible for Y iff it is legitimate to make the moral judgments involved in praising or blaming X for Y.
> 4. If determinism obtains, then every property that a person bears which might enter into the moral judgments involved in praise and blame holds of that person by luck.
> 5. So if determinism obtains, then if anyone is morally responsible, then moral luck exists.
> 6. So if determinism obtains no one is a morally responsible agent.

This argument makes no explicit reference to ability or control, but of course no one should take me at my word that it does not implicitly rely on those concepts. In particular premise four needs to be spelled out in a way that makes clear that it does not include a commitment to controversial claims about control and ability. Before I show that, however, I need to first discuss luck more generally, and another luck argument against compatibilism in particular.

1. Determinism and Luck

1.1 Levy's Luck Argument

Neil Levy has presented luck arguments against both compatibilism and libertarianism. According to Levy "luck is a function of three factors: significance, chance and control."[161] This general characterization of luck describes both constitutive luck and the various forms of non-constitutive luck, for which Levy gives the following accounts:

> **NCLuck**: E is a matter of non-constitutive luck iff E occurs in the actual world at time t*, but it fails to occur in a large proportion of possible worlds obtainable by making no more than a small change to the actual world at time t, where t is just prior to t*, and the agent lacks direct control over E's occurrence, and E is significant.
>
> **CLuck**: E is a matter of constitutive luck iff E is a trait that varies in human experience, the agent lacks direct control over whether or not E's is true of her, and E is significant.[162]

The reason for two different accounts is that constitutive luck, or luck in "the kind of person you are, where this is not just a question of what you deliberately do, but of your inclinations, capacities, and temperament,"[163] might not count as lucky on NCLuck, and not just because of the name. If it is the case that for agent X to count as the same person across possible worlds, that X have all the inclinations, capacities and temperament, then none of these things will hold of X as a matter of luck, according to NCLuck. Levy's particular worry is that to the extent that these inclinations, capacities and temperament are genetically determined, they will count as essential, for the same reasons that some have thought that one has one's genes essentially. So Levy gives us a separate account of constitutive luck that is supposed to be importantly similar to the primary definition of luck. It differs only in the gloss given to E being chancy. In NCLuck E is chancy because in some suitably large set of nearby possible worlds, E does not occur, while in CLuck E is lucky because it is a

[161] Levy (2009a)
[162] See Levy (2009a) and Levy (2009b)
[163] Nagel (1993) pg. 60

trait that varies widely in human experience.

This might seem like an ad hoc maneuver. That 'varying widely in human experience' and 'failing to obtain at a nearby possible world' are both things you could mean by the word 'chance' that does not imply that the two properties are related in any significant way. Perhaps 'chance' is not only ambiguous but ambiguous between meanings that have very little to do with one another. Could Levy avoid this potential problem by removing the concept of chance from the account of luck? In other words, would an account of luck just in terms of lack of control and significance be adequate? I do not think so.

1.1.1 Luck and Control

As Andrew Latus has pointed out, I have no control over whether it is the case that the sun will rise tomorrow, but it will not be a matter of good luck when it does.[164] It is not a matter of luck that oppositely charged particles are attracted to one another, nor is it a matter of luck, for anyone, that the theory of evolution is true. These kind of examples are precisely the reason that Levy gives for his hybrid account of luck rather than a lack of control account. I think that Levy moves too fast here. The lack of control account of luck has been too widely accepted for it to be likely that it could be brushed aside so easily.[165] So let us consider some alternatives to the claim imperiled by the aforementioned examples. From:

B: If X is significant for Y and outside of Y's control, then X is a matter of luck for Y.

One might move to:

> **B':** If X is significant for Y and outside of Y's control, and X is the kind of thing that people have a capacity to do, then X is a matter of luck for Y.

[164] Latus (2000)
[165] This conception of luck is appealed to in Nagel (1993). It is also appealed to in Statman (1993), Zimmerman (1987) and Greco (1995) among others.

Because causing the sun to rise is not within the capacities any human being has, it cannot serve as a counter-example to B'. B' still fails however, as seen in the following example:

> *Tiger Woods and Me*: It is clearly within a human being's capacities to hit a golf ball 400 yards. Tiger Woods can do that sometimes. Now I do not control the occurrence of the event 'Patrick hits a ball golf ball 400 yards'. I can want it to happen, and yet it will not. I can successfully will its non-occurrence, but that is not enough for control. Still it is not a matter of luck that I do not hit a ball 400 yards. I have not trained for it, and I have not learned the skills necessary to do it. This is a matter of a lack of skill, not of luck.

It is not a matter of luck that I fail to hit a golf ball 400 yards, yet it is something over which I have no control and something that is within the set of general human capabilities. In place of B' let me suggest:

B'' If X is outside of Y's control, and X is the kind of thing that Y has the capacity to do, then X is a matter of luck for Y.

I think B'' falls too, but it does not fall to the Tiger Woods and Me example nor to the sun rising example. That Tiger Woods has hit a ball 400 yards does not show that I have the capacity to hit the ball 400 yards.[166] The following example shows that B'' is false:

> *Michael Jordan and the Casino*: Michael Jordan has the capacity to dunk a basketball. But one night he plays and does not dunk once. Every time he tries he has to switch to the less athletically demanding lay-up at the last moment. What no one but Jordan and some guys in Vegas know is that Jordan was up gambling all night, and was drinking and snacking constantly while doing so. Because of that Jordan's dunking is not responsive to Jordan's wanting to dunk, so plausibly he is not in control of whether he dunks.

[166] I imagine that some might be uncomfortable talking about whether individual people have capacities. I do not think the term capability sounds nearly so odd in this context, and so I encourage the reader to substitute 'capability' for 'capacity' if she wishes to.

It is not bad luck that Jordan can't dunk. It is his own fault he can't. But no one would question that Michael Jordan is capable of dunking. He just can't tonight, no matter how much he tries, and so, plausibly, he doesn't control whether he dunks or not. The way typically to fix such problems is to put in a tracking clause. So:

> **B'''** If X is outside of Y's control and that lack of control was not caused by some action which was both in Y's control and could reasonably be expected to cause that lack of control, and X something Y has a capacity to do, then X is a matter of luck for Y.

B''' is the best version of **B**, but I think it still fails. It is overbroad. Too many failures that are not matters of bad luck would be classified as such by **B'''**. Take, for instance:

> *Prosecution and Defense*: In a trial where the guilt or innocence of the defendant is not obvious, both the prosecution and defense do all they can, within the limits of the law, to win. Neither fails to pursue any legal option that can reasonably be expected to influence the outcome of the trial. Because the guilt or innocence of the defendant is not clear, and because the judge gave firm instructions to the jury not to convict unless the prosecution proved their case beyond a reasonable doubt, the prosecution loses the case.

Because the prosecution lost both due to the facts of the case being as they were and general features of the legal system, it seems pretty clear to me that the prosecution did not lose as a matter of luck. According to **B'''** they did. It was certainly not in the prosecution's control whether they won. It was also not because of some failure on the prosecution's part that they lost. In fact the problem **B'''** has with *Prosecution and Defense* will show up in any case of fair competition. If both competitors try their best, one is not going to win, her failure to win will not be under their control, and its not being under her control would not be attributable to some action of theirs. But not every result of fair competition is a matter of luck.

It does not seem that an account of luck based on the lack of control is going to

be adequate. As we saw if we add the presence of chance as a necessary condition the consequence is the seemingly ad hoc distinct accounts of constitutive and non-constitutive luck. Of course, the ad hoc nature of the account is only a problem if we assume that there must be some clearly unified and simple account of luck available. Should we assume that there is such an account available? I don't think so. Certainly not every concept can be given such an account, and there does not seem to be any special reason to insist that luck must be. In fact I will argue later that the real problem with Levy's account of luck is not that it is ad hoc, but rather that it is too simplistic. Luck is a complex phenomenon and Levy's account fails to capture that complexity.

1.1.2 From Luck to Incompatibilism

Before presenting that criticism, let me put Levy's argument fully on the table. First, while Levy's argument itself is directed only at a subset of compatibilists, it is part of a larger argument that has as its target compatibilism generally. Levy claims that his argument aims to show only that historical accounts of compatibilism face a problem of present luck. 'Present luck' is a term taken from Mele[167] that refers to luck in and just around the moment of action. It is distinguished from remote luck, which is luck significantly before the moment of action. Why do historical accounts of compatibilism face a present luck problem, according to Levy? Historical accounts of compatibilism, as the name suggests, deny that whether or not a person is free and responsible depends only on the non-historical facts true of the agent at the time of action. Non-historical compatibilists tend to think that the necessary and sufficient conditions for freedom and responsibility are exhausted by facts about the agent's psychological makeup at the time of action, usually whether the agent endorses her course of action in the right way. Historical compatibilists demand that the

[167] Mele (2006)

psychological makeup have been produced in a certain way, that it have a certain kind of history.

What kind of history is required? Levy rightly points out that a popular answer to this question is that the history must be characterized by the agent taking responsibility for not only her actions, but also the psychological grounds of her actions. The problem that Levy presents is that because taking responsibility is an action, it stands in need of an explanation, and in all likelihood it is going to be explained by the very psychological features that it is supposed to ground. So, for example, if Alex has a tendency to be lazy, and he takes responsibility for that laziness, but only does so because of the complacency with his laziness that is itself a product of his laziness, it is hard to see how his taking responsibility can make him responsible for his laziness. It looks as if Alex could only be responsible for taking responsibility in this case, if he was already responsible for his laziness. Now, it will of course be rare for the trait which we are deciding about when we take responsibility to be the trait that explains our deciding that way. But if all our traits and dispositions, all our psychological makeup in fact, are such that we cannot be responsible for them unless we take responsibility for them, then, it looks like, we cannot effectively take responsibility for them unless our taking responsibility for them is not determined by those traits and dispositions.

The problem that faces historical compatibilists is similar to one faced by event-causal libertarians. According to event causal libertarians our decisions, or some important subset of them, are undetermined events. The problem that this raises is that it seems as though the action is just a matter of luck, because there is no explanation of why it happens rather than some other action. This is the problem of present luck that faces event-causal libertarians, and Levy's argument is that the very same problem faces historical compatibilists.

Now on the face of it this argument seems rather easy to get around. One might just affirm non-historical compatibilism and avoid the problem of present luck entirely. Alternatively, one might affirm historical compatibilism but not think that the relevant historical feature is one of 'having taken responsibility'. Gary Watson, for instance has affirmed a historical view on which all that is added to a standard non-historical view are the requirements that one's psychological makeup not have been determined purely by acculturation.[168] Either by affirming non-historical compatibilism or by affirming a version of historical compatibilism which does not trace responsibility for the agent's psychological makeup to actions the agent herself has performed, it looks as though one can get around Levy's argument.

That appearance would be deceiving however. What one can do is get around Levy's present luck argument. One would not thereby get around the overall luck argument against compatibilism that Levy has presented. The reason why is that Levy's argument essentially consists of two dilemmas, one embedded within the other. The dilemma we are already familiar with is the one faced by historical compatibilists who affirm a 'taking responsibility' condition on moral responsibility for traits and dispositions. The dilemma there is that either these compatibilists must accept that the actions involved in taking responsibility are explained by the traits and dispositions they are about, in which case they cannot be the ground of moral responsibility, or they are not explained by those traits and dispositions, in which case it seems as though it is just a matter of luck whether the person took responsibility or not.

The second, broader dilemma is one faced by compatibilists generally. It is a dilemma on which, on the one hand you can adopt historical compatibilism and face

[168] Watson (1975) It is not entirely clear what Watson means here. I take it that he has in mind ruling out people being responsible for actions which are explained by psychological traits where are themselves simply explained by a process of education and upbringing which determined the person to have those psychological traits in some highly abnormal way, perhaps a way which completely bypasses the person's rational faculties or perhaps one that involved significant harm to the person.

the first dilemma, or you can adopt non-historical compatibilism, in which case you face the problem of constitutive luck. Why do non-historical compatibilists face a problem of constitutive luck? Levy is not entirely clear on this point. He says that non-historical compatibilists face the problem of manipulation, and that the problem of manipulation is equivalent to the problem of constitutive luck. How to interpret this?

The manipulation problem that faces non-historical compatibilism is basically that it seems as if there is no principled way to distinguish between an agent who meets all the conditions for responsibility endorsed by non-historical compatibilism because she was manipulated into meeting those conditions, and an agent who meets those conditions in some more normal way.[169] This is a significant problem for compatibilists. Manipulation is plausibly a defeater of moral responsibility claims, and if there is no principled difference to be drawn between the manipulation cases and cases of non-historical compatibilist agency, then non-historical compatibilism either fails or has to claim that manipulated agents can be morally responsible. It is not clear which is the worse alternative. What is also not clear is how the manipulation problem could be equivalent to a luck problem. A luck argument would have as its conclusion that a person's actions only happen by luck. The manipulation argument has as its conclusion that compatibilist accounts of moral responsibility are overbroad, classifying some cases of manipulation as cases where the agent is morally responsible for what she does. These don't seem like equivalent problems, nor does it seem that the conclusion of a luck argument is implied by the conclusion of a manipulation argument.

There is a way to make sense of Levy's claim though. If it were the case that everything a manipulated person does she does just by luck, then if someone could be

[169] See Kane (2008) and Pereboom (2001)

manipulated into meeting the non-historical compatibilist conditions on responsibility, then someone could meet those conditions while her action was just a matter of luck. So non-historical accounts of compatibilism would then imply that people could be responsible for what is a matter of luck, and that is generally agreed to be impossible. But is it the case that manipulated actions are just matters of luck? Looked at from the point of view of the manipulator it seems not. After all the mad scientist controlling Jane via mad scientist technical gear will not think it was a matter of luck that she robbed the bank and brought the money back the mad scientist's lair. The mad scientist will say 'Its not luck that she came here, I spent years developing this technology so that with the push of a button I would be enriched by the activities of others.' From Jane's perspective it will seem like very bad luck that she robbed the bank. From her perspective it will seem like once she got under the control of the mad scientist it is just a matter of luck what she ends up doing. In this case she had the bad luck that the mad scientist wanted some money. It would have been good luck if the mad scientist had wanted to support his friend by sending his manipulated subjects to see the friend's Broadway play. Jane's actions are all determined by the whims of someone she has never met and can have no influence over.

How do we resolve the conflicting impressions we get from these two perspectives? We don't. What we should do is introduce more precision to this discussion of luck. I suggest that we stop taking there to be a simple property 'being a matter of luck'. When we say that X is a matter of luck I suggest we take that claim as elliptical for 'X is a matter of luck for everyone.'[170] The flip side of this suggestion is that we take it that claims about luck have the form 'X is a matter of luck for Y' where X is some state of affairs and Y is something that can be lucky. What things can be lucky? Humans clearly can. I think animals clearly can. After all the deer is lucky

[170] With some suitably contextually restricted quantifier.

when the hunter is so inept that he drops his gun while bringing it up to aim. I think plants might be lucky as well. Perhaps it is lucky for the oak that its bark was thick enough to withstand the small forest fire, and bad luck for the ash that it could not. Can rocks be lucky? I doubt it. Numbers and other abstracta certainly can't be. I don't think that much of importance hinges on specifying the class of objects which can be bearers of luck, because all we are interested in here is the class of objects that can be morally responsible, which is clearly entirely contained within the class of objects that can be bearers of luck.

With this revision in hand I think we can safely say that Jane's robbing the bank was a matter of luck for Jane but not for the mad scientist. This is a result that suits Levy fine, because Jane did not have control over the bank robbing and yet the mad scientist did. In fact any manipulated action that was suitably chancy would be a matter of luck for the victim of manipulation, according to Levy, because of her lack of control. Would manipulated actions be suitably chancy? Not all would be. If the manipulator suffered from a pair of mental illnesses, one which produced the compulsion in him to acquire other people's things, and the other which produced in him a compulsion not to be directly involved in crimes, then, along with his evil scientists manipulation know-how, this would mean that in all the nearby possible worlds the manipulator would be manipulating Jane.[171]

But it seems that something has gone wrong here. Why should it matter, when we are figuring out whether something is a matter of luck for Jane, that the evil scientist has these particular compulsions? His having them, rather than some temporary and chancy desire to have others rob banks for him is nothing to Jane. So I suggest another revision to Levy's view, in the same spirit as the first. It is to replace

[171] Jane being the person who is regularly vulnerable to being implanted with the mad scientist's technology. Suppose Jane was assigned, years ago, to be the mad scientist's cleaning lady.

the requirement of chanciness with the requirement of something's being chancy for an agent. The problem is that it is not obvious how we alter Levy's account of chance to accommodate something being a matter of chance for one person and not for another. After all, if something fails happens at all the nearby possible worlds but does happen in the actual world, it doesn't matter who you are, that thing is just a matter of chance on Levy's account. My suggested revision to Levy's account is to add to his account of chance that if, for all a person knows or can be expected to know, X is a matter of chance, according to Levy's original account, then X is a matter of chance for that person.

Why should we accept this intrusion of epistemic notions into our account of chance? Because the intrusion is supported by some pretty common intuitions:

> Astronomer's Wager: Jon, Jane and Pat are astronomers of reasonable financial means. They have been working for years on finding out information about a particular planet 400 light years away. They know its mass, its location relative to other planetary bodies in its solar system, and that it has an atmosphere. What they do not know is what the atmosphere is composed of. They know that there are three options and that each option for the chemical composition of the atmosphere corresponds to a color that the atmosphere will have. So they decide to take bets on what the color of the atmosphere is. They will find out what color the atmosphere is the day after the bet is made. But they have no information about what the chemical composition of the atmosphere is, let us say. They have only just recently invented a device that allows them to gather information, and it will not go online until the day after the betting. So when they pick which color to go with, they do so in the absence of all information. Let us suppose that Jon, Jane and Pat all have, and have had most of their lives, a liking to one of the possible colors, and not the others. So in the absence of information they are just going to pick their favorite colors. Let us also suppose that their favorite colors are all different, so that each can pick their favorite color without conflicting with someone else. Jon picks blue. The next day the information comes in, and the planet has a blue atmosphere.

I think it is a matter of luck for Jon that he wins the bet. But his winning the bet is not chancy according to Levy. Given the distance of the planet, what information they

were going to get about its atmosphere was set hundreds of years before the bet, so no small change just prior to the bet, or to the information arriving could have changed the color of the planet. All three astronomers have had their favorite colors for years, and so again no small change could have produced a different arrangement of bets. But clearly the fact that they picked in complete ignorance makes it a matter of luck for Jon that he wins.

So I think we need an epistemic account of chance, not a metaphysical account of chance, if we are going to make Levy's argument work.

> **EpNCLuck**: E is a matter of non-constitutive luck for Y iff E occurs in the actual world at time t*, but for all Y knows at t, it fails to occur in a large proportion of possible worlds obtainable by making no more than a small change to the actual world at time t, where t is just prior to t*, and the agent lacks direct control over E's occurrence, and E is significant.

Does this suggested revision force upon Levy something he would find objectionably radical departure from his original account? It would not amount to a radical departure from the claim that metaphysical chance is sufficient for luck, because metaphysical chance is sufficient for epistemic chance, which on this account is sufficient for luck. Metaphysical chance is not necessary for luck on this account, however. Something could fail to be metaphysically chancy but still turn out to be lucky on **EpNCLuck**. But I don't think Levy should find that objectionable. That means more things can be counted as lucky than could on **NCLuck**, and since Levy's goal is to argue that everything we do is a matter of luck, he could hardly find reason to complain.

So with this new version of Levy's account of luck, I think we can safely assume that all manipulated actions are a matter of luck for the person being manipulated, and so that if the standard manipulation arguments are sound, then

people are lucky with respect to their constitutive traits and dispositions. And that result, combined with his present luck objection to historical compatibilists, means that Levy has a general argument from premises about luck against compatibilism generally.

1.2 Against Levy's Luck Argument Against Compatibilism

There are two reasons to reject Levy's luck argument and look for another. One, the account of luck that Levy gives is false. Two, the account of luck Levy gives deprives his argument of any force not already had by other incompatibilist arguments, because of controversies over the nature of control. The second problem is easier to appreciate so I will start there. To see how controversies about the nature of control matter, that the proper account of 'control' as it appears in Levy's account of luck, is one on which having control over one's actions is possible if determinism is true. Since the absence of control is a necessary condition of luck, on Levy's account of luck, it would mean that some actions, the ones that are under our control in a compatibilist sense, are not such that it is a matter of luck for us that we perform them. Why is this so? Remember that Levy's dilemma depends on the assumption that because our constitutive traits and dispositions are matters of luck for us[172], actions wholly explained by those traits and dispositions are matters of luck. This assumption is false if we assume a compatibilist account of control along with Levy's account. That some action is explained wholly by the presence of dispositions that are outside of an agent's control does not imply that the action is outside of the agent's control, according to standard compatibilist accounts of control. So even if it is a matter of luck which constitutive traits you have, it is not thereby a matter of luck what actions you perform, even if those actions are explained entirely by those constitutive traits, if

[172] No compatibilist account of control implies that we have control over those traits and dispositions, so this claim is not imperiled by the hypothetical acceptance of a compatibilist account of control.

we assume a compatibilist account of control.

But if Levy demands that we accept an incompatibilist account of control, then it turns out that luck doesn't play an important role. If the argument only works because of the assumption of an incompatibilist account of control, then why not think that it is typical incompatibilist reasons that show that compatibilism is false, rather than these new impressive seeming reasons having to do with luck? Put another way, the only way for Levy's argument to work is for standard incompatibilist arguments to also work, and that raises the suspicion that it is standard incompatibilist intuitions that are expressed by, for example, the Consequence Argument, doing the work for Levy. A different version of the same problem reveals itself when you consider how dependent Levy's argument is on the claim that there is no principled difference to be drawn between non-historical compatibilist free agency and manipulated agency. Levy's argument is sound only if manipulation arguments are sound. If there is some principled difference between cases of manipulation and non-historical compatibilist accounts of agency, then Levy's argument is unsound.

Now those problems might seem somewhat trivial. After all Levy can be thought of as engaging in the task of supplementing or strengthening incompatibilist arguments. The fact that Levy's argument is sound only if other incompatibilist arguments are sound is not relevant to that task. That Levy's argument is not independent does not stop it from having some function in the debate. It would be interesting, after all, to know that there was a luck argument against compatibilism, since there is often thought to be a luck argument against libertarianism. A luck argument against compatibilism would provide some reason to think that the skeptical worries about freedom and moral responsibility can be unified under the description of 'problems of luck'. Even if Levy does not provide independent reasons to reject compatibilism, he could reasonably claim to have provided a new way to think about

the problems that face a free will defender.

I don't think Levy has done that much though, because I think the account of luck he has offered fails. I do not think chance plus lack of control plus significance is either necessary or sufficient for luck. In particular I want to deny that chance is a necessary condition of luck, and deny that chance, the lack of control, and significance are together sufficient for luck. First to chance:

> *Lottery and Near Cheating*: Tom plays the lottery every week, picking the same set of numbers. After years of failure Tom's numbers one day come through for him, and he wins millions. Unbeknownst to Tom, his friend Bill, wanting to see Tom finally win, hacked into the lottery computer system just that week, intending to switch the winning numbers to Tom's numbers. But, Bill finds out, Tom has actually won on his own. Had Tom not won on his own though, Bill, a skilled hacker, would have successfully changed the numbers to make Tom win.

I take it that no small change, prior to Tom's winning would make it the case that Tom won. So it is not a matter of objective chance that Tom won, but it is, I think, a matter of luck for Tom that he won. This is not surprising given the lessons learned from *Astronomer's Wager*. There we saw that objective chance was not necessary for luck, but that epistemic chance might be. And if we are considering Tom's being lucky in winning, we can see that there is no counterexample to the epistemic chance condition on luck here. But what about Bill? It is very important for Bill that Tom do well. Tom's winning the lottery, therefore, means that a deep desire of Bill's is satisfied, and let us say no conditions are operative which prevent the fulfillment of this desire from constituting a good for Bill. Is it a matter of luck for Bill that Tom won the lottery?

I think it is. If I am right then *Lottery and Near Cheating* constitutes a counter-example to the epistemic chance condition, because after all, given Bill's knowledge of his own hacking skills and of the nature of the computer system he was

to hack into, there was no chance that Tom was not going to win prior to Tom's winning or prior to Bill's actually hacking into the system. What makes that true is Bill's intentions to intervene if the numbers chosen did not match Tom's lottery numbers. But since those intentions were impotent in the actual case, it still seems like a matter of luck, for Bill, that Tom won.

Also, neither objective nor epistemic chance are sufficient for luck. This does not follow from *Lottery and Near Cheating* of course, but it can be made plausible by considering the intuitive ubiquity of chance in our lives, and how unintuitive it is to think that luck is similarly ubiquitous. Consider the following case:

> *Lazy Thief*: Jon, Thomas and Alexandra have plotted to rob a bank. These three are very sophisticated thieves and they have prepared countermeasures for every eventuality. Jon has the task of making it impossible for them to be tracked after the robbery has been carried off. Jon is lazy however, and neglects to cover the outfits they will be wearing with a substance which would make it impossible for police dogs to track them. They rob the bank and are pursued by dogs. Given that Jon neglected to properly prepare, it is possible for the dogs to find them. It is however, very unlikely they will do so, as the three thieves take normal precautions against it (crossing rivers and the like). The dogs do manage the track them however, against the odds.

I don't think it is a matter of luck for Jon that the three are caught, despite the fact that it is clearly a matter of chance that they are caught. The fact that had Jon taken adequate precautions, the three would not have been caught, makes it the case that it is not a matter of luck that they are caught, despite the fact that even without the precautions it was still unlikely they would be caught.

So chance is neither necessary nor sufficient for luck. It has been argued, by Jennifer Lackey in particular, that lack of control is neither necessary nor sufficient for luck. Levy thinks that lack of control is necessary but not sufficient for luck, and here I agree with him. But what about the question of whether chance, plus lack of control plus significance is sufficient for luck? So far I have presented no counter-examples

to that. If it were the case that *Lazy Thief* involved a lack of control, on Jon's part, about whether the thieves were caught, then it would constitute such an example. It seems, though, that it is precisely because Jon had it in his control whether it was impossible for them to be caught that it is not a matter of luck for Jon that they were caught. Looks may be deceiving here, because it is certainly not in Jon's control whether they were caught. After all, their being caught was not something Jon intended, and it was something Jon took steps to prevent, though unfortunately for his fellow thieves he did not take all the relevant steps. Jon did not bring it about that they were caught, he simply omitted an action that would have made it impossible for them to be caught. This case produces a dilemma for Levy, because the fact over which Jon had control (its being possible that they would be caught by the dogs) is not chancy, and the fact that is chancy (their being caught by the dogs) is not something over which Jon exercised control. So Levy cannot use the strategy of relocating the luck or absence of luck.

1.3 A Different Strategy

Levy's luck argument against compatibilism is the most developed version in the literature. As we saw it both relies on a suspect account of luck and wouldn't, even if Levy's account of luck were correct, provide a reason to reject compatibilism unless standard incompatibilist arguments could be shown to do so. As I said, I am interested in showing that previous forms of compatibilism fail to respond to the central incompatibilist insight, and the way to show that is by presenting a recognizably incompatibilist argument that does not rely on any of the contested notions found in standard incompatibilist arguments. I want to avoid talk of control, not build a substantive account of it into an account of luck.

So if I can't rely on an account of luck that includes the concept of control, what account of luck can I rely on? I don't think there is an account of luck that is

adequate. Levy's account comes closest of existing accounts to getting things right, and even it undersells the complexity of the concept. I know of no way to improve on Levy's account of luck that would not amount to ad hoc additions to the account that deal with the counterexamples I raised. Does this mean that it is impossible for me to offer a luck argument against compatibilism? No, for one does not need a fully worked out account of luck in order to be reasonably sure of certain claims about luck.

I will now consider two ways of arguing for premise (4) of the Moral Luck Argument, which says that if determinism obtains then every property that a person bears which might enter into the moral judgments involved in praise and blame holds of that person by luck. Each will be supported not by a full account of luck, but by plausible claims about luck.

1.3.1 Transfer of Luck

The first way of arguing for (4) will look very familiar, as it is very similar to the Consequence Argument and Direct Arguments for incompatibilism. The Consequence Argument and Direct Argument both appeal to transfer principles. The Direct Argument appeals to the transfer of lack of responsibility principle, which says that if X is sufficient for Y, and you are not responsible for X then you are not responsible for Y. The varied versions of the Consequence Argument also appeal to transfer principles. One, for example, appeals to the principle that if you do not have a choice about X and X is sufficient for Y, then you do not have a choice about Y. Another appeals to the principle if that you do not control X and X is sufficient for Y, then you do not control Y.

These transfer principles have all been subjected to significant compatibilist scrutiny, and it is easy to see why. If determinism obtains then the deep past along with the laws of nature are sufficient for every fact which might ground an attribution of moral responsibility. All of our thoughts, actions, and character traits are the

consequences of facts about the deep past and the fundamental structure of the universe, two things that we are not responsible for, we do not control, and that we do not have a choice about. If the transfer principles mentioned above are valid principles of inference, then determinism implies that we are not morally responsible for, do not have a choice about, and do not have under our control anything at all.

I do not want to take a stand on the validity of these inference principles. What I want to do is take a quick look at what an argument that employs the following transfer principle can do:

> **Transfer of Luck**: If X is a matter of luck for p, and X is sufficient for Y, then Y is a matter of luck for p.

Transfer of Luck shares the form of the other transfer principles, with the difference being that it is about luck rather than control, choice or responsibility. The argument it suggests shares the form of the Consequence and Direct Arguments, but is about luck rather than control, choice or responsibility. That argument is:

> i. If determinism obtains then a proposition expressing the state of the world at some time, T, deep in the past (P_0) taken together with a proposition expressing the laws of nature (L) imply every proposition expressing the state of the world at a time later than T, including the proposition expressing the state of the world at the present time. [$(P_0 \& L) \rightarrow P$]
> ii. For any agent at the present time, T_p, P_0 is a matter of luck.
> iii. For all agents L is a matter of luck.
> iv. So for any agent at T_p (P_0 & L) is a matter of luck.
> v. So, by **Transfer of Luck**, P is a matter of luck for every agent at T_p.

I am going to grant that (4) follows from (2) and (3), even though the inference is invalid.[173] The problem does not lie there and I am sure that with some work someone could justify the inference for a property like 'is a matter of luck for'.[174] What

[173] *A*X & *A*Y does not imply *A*(X & Y). Suppose *A* was the property 'is not a conjunction' or 'has no parts'.

[174] Some reason to think so is that the excuse 'X was just a matter of luck' does not fail just because some part of X was not a matter of luck. When a gambler wins at roulette, that is a matter of luck, even

concerns me are premises (2) and (3), as well as **Transfer of Luck** itself. Do we have reason to accept any of these claims?

I think we have reason to accept (2) and (3). I think that a case can be made that the laws of nature are a matter of luck. As an example of a law of nature that might hold by luck I would point to the relationship between the gravitational constant and nuclear force. If the ratio were different life could not exist, or at least not life as we know it. Every living thing is composed of carbon and oxygen, among other elements. These elements are the product of fusion which took place in the core of a star. The ratio between nuclear force and gravity determines how long a star can exist. Gravity exerts pressure keeping a star from exploding, while nuclear force exerts pressure pushing a star to explode. Eventually all stars do explode, but given the amount of time they have before that event, carbon and oxygen can be created in the core of the star. If the ratio of gravitational force to nuclear force were different enough from what the ratio is now, stars would not survive long enough to produce the elements necessary for life. We are all lucky, it seems, that gravitational force bears the relationship it does to nuclear force.

There are a variety of reasons one might question this example though. Some physicists, like Martin Rhees, have claimed that for every possible set of natural laws there is a corresponding actual universe and that because of this fact it is not a matter of luck that the natural laws are so 'fine-tuned' for supporting life.[175] One might wonder why this claim about a multiplicity of universes would, if it were true, imply that the laws of nature being as they are is not a matter of luck. One might think that this claim shows that it is not a matter of chance that a universe with just our laws

---

though part of what made it the case that he won at roulette, that he decided to play roulette that night, was not a matter of luck. If someone's house gets hit by a tornado, the insurance company could not point to the fact that it was not a matter of luck that they bought a house at that location in order to show that it was not a matter of bad luck that the house got hit by a tornado.

[175] Rhees (2000)

exists, since for every possible set of laws a universe exists with those laws. But we cannot infer from the fact that X is not a matter of chance, that X is not a matter of luck. As we saw in *Lottery and Near Cheating* chance is not a necessary condition for luck. Perhaps what is meant is that it is metaphysically necessary that there be a universe with laws like ours, and also necessary that if human being exist we exist in that universe, and because of this 'Our universe having the laws it has' cannot be something we are lucky or unlucky with respect to. Can we be lucky that metaphysically necessary facts obtain, or that necessary truths are true? If **EpNCluck** is correct then we can be, because it could be beyond our ability to know that something is metaphysically necessary. Presumably the existence of other concrete actual worlds is just the kind of thing that is beyond our ability to know.

In ordinary discourse one is much more likely to find people treating events in the past as matters of luck than treating the laws of nature as matters of luck. To give a personal example, my very existence seems to hinge on a past event which it seems plausible, to me at least, to treat as a matter of luck. My paternal great-great grandparents knew each other as children. They were separated by the violence and turmoil that plagued central Europe in the late 19th century and only ever got married because of a lucky meeting at a train station. My great-great grandfather was scheduled to depart from that train station around the same time my great-great grandmother arrived. He recognized her and they struck up a conversation, leading him to miss his train, leading to them spending more time together that day, leading to them starting a relationship, leading, through many more steps, to me. Everyone on my father's side of the family treats it as a matter of good luck that our recent ancestors happened to see each other that day at the train station.

Historical examples of past events being treated as matters of luck abound. It was a matter of bad luck for Native Americans that they had not developed resistances

to common Eurasian diseases prior to the arrival of Spanish galleons. It was good luck for the residents of Kokura and bad luck for the citizens of Nagasaki that clouds covered Kokura on the day that the atomic bomb was to be dropped on Kokura, causing the Americans to instead destroy Nagasaki. It also seems pretty clear that, for descendants of those who suffered the initial good or bad luck that these events are still matters of good or bad luck.

Even if one does not accept the examples offered, another argument that events in the deep past can be matters of luck starts with the premise that events in the recent past can be matters of luck. It is a matter of bad luck that the economy crashed when it did for everyone who found themselves without a job at the time. If one wished to maintain that events in the deep past were not matters of luck, one would have to think that at some point the event 'the economy crashing' stops being a matter of bad luck and starts not being a matter of luck at all. At what point would that happen and what would explain this change?

I think that premises (2) and (3) are more plausible than their denials. If **Transfer of Luck** is a valid inference principle, then if (2) and (3) are true, then if determinism obtains everything is a matter of luck. Given standard views about the relationship between luck and moral responsibility, this would amount to an argument for the incompatibilism of determinism and moral responsibility. But **Transfer of Luck** is not a valid inference principle. Imagine a gambler who has been down on his luck. He has the quite common but also false belief that his losing more often than not at gambling means he is suffering from a case of bad luck, and therefore that he will fail in other enterprises as well. This gambler is impulsive however. The moment he feels his luck has changed he will, though he does not know it is so ahead of time, be filled with such confidence that he will immediately find the woman he loves and ask her to marry him. When the gambler wins at roulette, he collects his winnings, leaves

the casino, drives to his girlfriend's house, and asks her to marry him. It is certainly a matter of luck that he won at roulette. And his winning at roulette caused him to ask his girlfriend to marry him.[176] But it doesn't seem, to me anyway, that his proposing to his girlfriend is not a matter of luck. After all, he proposed to his girlfriend because he loved her and because he thought that decisions taken at that time were bound to go well for him. That he was caused to believe that by a lucky event does not make what follows from that belief a matter of luck.[177]

It is for this reason that I think the first strategy will not work. But one need not endorse a 'luck transfer' principle in order to make use of the central insight of the Consequence Argument, that determinism imperils moral responsibility because it says that how we behave is the consequence of events in the deep past and the laws of nature. What I will argue is that our actions are a matter of luck, because they were 'in the cards' from long before our birth. The argument for this claim proceeds, like all the arguments so far, by appeal to examples. I will start with examples that are uncontroversially cases of luck and try, through successive small alterations, to motivate the claim that our actions are not importantly different from uncontroversial cases of luck if determinism is true. I am left with this roundabout method of arguing that determinism implies that everything is a matter of luck because I have rejected all the current accounts of luck, and because I have rejected the **Transfer of Luck** principle discussed above. I cannot argue that determinism implies luck from some general account of what makes something a matter of luck, because I do not take myself or anyone else to be in possession of such a general account. And I cannot let

[176] As with the causal relationship between the past and the present, there are background conditions in place in the gambler story that make it the case that his winning at roulette is the cause of his proposal. Winning at roulette is not a logically sufficient condition for the proposal.

[177] What might be a lucky event is his proposing to his girlfriend around the time he actually proposed. After all if he had only started winning a week later, he wouldn't have proposed to her for another week. But from the fact that it is a matter of luck that X happened at time T, it does not follow that it is a matter of luck that X happened. It is a matter of luck that the roulette wheel hits 7 when it does, but it is not a matter of luck that the roulette wheel hits 7 at all. It is bound to happen sooner or later.

the argument rest on the intuitiveness of the claims that the past and laws of nature are matters of luck both because it is not clearly intuitive that they are, and because even if they are luck does not transfer along lines of logical or casual sufficiency.

1.4 A Better Attempt at a Demonstration

I want to start with an uncontroversial case of luck. I take the following case of gambling to be just such a case:

> *Equal Poker:* Jack plays poker for a $1,000 pot. There is nothing odd about the deal, it is a standard poker dealer doing his job as he ought to. The only odd thing about the players is that, while they all know only their cards, they are all equally skilled at calculating the likelihoods of other hands being present in the game given their knowledge of their own hands, at knowing how and when to bluff, etc.. Jack wins $1,000.

It strikes me as uncontroversial that Jack's winning the money is a matter of luck. One might say it is partly a matter of skill, but the fact that they are all of equal ability speaks against that. Jack's win does not express or manifest any skill not possessed by the other players, and let us suppose that no one at the table made a mistake uncharacteristic of a player of her skill. I take it that this win is just a matter of luck. Now consider:

> *High Stakes Equal Poker*: In this situation the same factors obtain as in *Equal Poker* except that the pot has changed. Instead of being a game for money, the pot determines which opportunities for careers, social standing, and living environment the children of those at the table will have. Jack wins, earning for his daughter the chance for a great job, high social standing, and great material conditions throughout her life.

I take it that Jack's daughter having the life she has is just as much a matter of luck as Jack's winning $1,000. Changing the stakes does not seem as though it can make a game of luck something other than a game of luck. Now consider:

> *Absurdly High Stakes Equal Poker*: The situation here is the same as in *High Stakes Equal Poker* except that here every aspect of the lives of the children of the players is determined by who wins, who comes in

> second, etc. Every hard fact concerning the children of the players is set by who wins the pot.[178] Jack wins, wining for his daughter the same great life, but with every last detail set completely.[179]

Again only the stakes have changed, and so it does not seem as though there could be any principled move distinguishing *High Stakes Equal Poker* and *Absurdly High Stakes Equal Poker*. The next change will not be one of changing the stakes, but of changing the players:

> *Computerized Absurdly High Stakes Equal Poker*: Everything is the same as *Absurdly High Stakes Equal Poker* except that it is no longer Jack and other parents playing for the lives of their children. Rather, it is sophisticated poker playing computers that are playing out the hand. All the computers have equal computational abilities, mirroring the similarity of skill in *Equal Poker*. Jack's daughter get's the same life she got in *Absurdly High Stakes Equal Poker*.

It is again hard to see what could make it the case that we should judge *Computerized Absurdly High Stakes Equal Poker* differently than *Absurdly High Stakes Equal Poker*. It doesn't seem as though it should matter that it is computers, rather than people playing. Contests can be matters of luck even though no people are competing. Horse races are matters of luck, plausibly. The next step in this train of examples is:

> *Lonely Computerized Absurdly High Stakes Equal Poker*: This example is exactly the same as *Computerized Absurdly High Stakes Equal Poker* except that there is only one computer playing. It is playing against itself in the same way that a computer can simulate a chess or baseball game. The programs of the competing computers are simply brought together into one machine in this example. Once the cards are dealt the computer follows out its simulation program, a program which exactly mirrors what occurs in *Computerized Absurdly*

[178] By 'hard fact' I mean those facts about the world at one time that do not logically imply facts about the world at some earlier or later time. I mean to exclude all facts that van Inwagen meant to exclude from his specification of the proposition expressing the state of the world in the deep past in his version of the Consequence Argument.

[179] This case would not make sense if it were the case that Jack's daughter already existed, for if she did there would already be some relevant facts true of her. So this example should be treated as though Jack is playing for his unborn, and probably unconceived, child. As for how it could be that all the facts about her life are set before she is born, I leave it to the imagination of the reader to provide their own favorite philosophical bogey man to do the trick; an evil demon will work.

> *High Stakes Equal Poker*. Jack's daughter gets the same life she got in *Absurdly High Stakes Equal Poker*.

If it doesn't matter that we replaced the people with computers, it hardly seems to matter that we combined the computers. It is no less a competition than it was in *Computerized Absurdly High Stakes Equal Poker*, and presumably that was no less a competition than *Absurdly High Stakes Equal Poker*, or at least if it was, this made no difference to whether it was a matter of luck how Jack's daughter ended up. Time for the penultimate example:

> *Super Lonely Absurdly High Stakes Poker*: There is no computer player in this case. The cards are dealt and they determine by themselves how Jack's daughter's life ends up. After all a computer could have simply inspected the cards in *Lonely Computerized Absurdly High Stakes Equal Poker* and read off of them how each child's life would go. There was a mathematical formula that could describe the operations of the lonely computer, and that mathematical formula could just be used to ground the inference from the cards as they are dealt to the life that Jack's daughter will have.

It is hard to tell what the difference is between this case and *Lonely Computerized Absurdly High Stakes Equal Poker*. There are no players, and there is no competition, but this process will yield exactly the same outputs as the computer played games in the earlier examples, give the same initial deal. It is hard to figure out where the luck goes away in this story. But if it hasn't left us yet, then I do not see how we can avoid saying that in the next example how Jack's daughter's life goes should be considered a matter of luck:

> *Super Lonely Absurdly High Stakes Poker with a Cosmic Dealer*: There are no more cards being dealt, it is initial arrangements of energy and matter in the universe that take the cards' place, and the mathematical formula which read the result off the deal is replaced by a much more complex formula consisting of all the laws of nature and the laws of logic.[180] The initial input, fed into the formula, invariably

[180] Or if you like the laws of nature are part of the package of information that replaces the cards as inputs to the formula, and the formula is just the laws of logic.

> yields the same life for Jack's daughter. The only variances in Jack's daughter's life comes from varying the initial arrangement of energy and matter in the universe.

If luck has been preserved at every step in the iteration of examples, then I have my argument, because *Super Lonely Absurdly High Stakes Poker with a Cosmic Dealer* is really just the situation that obtains if determinism obtains, or so close to it that no one could hope to block the inference from the proposition that Jack's daughter's life is a matter of luck in *Super Lonely Absurdly High Stakes Poker with a Cosmic Dealer* to the proposition that Jack's daughter's life is a matter of luck if determinism obtains. And of course there was nothing special about Jack or his daughter in the example. This result would generalize. Every hard fact that concerns us would hold, for us, as a matter of luck if determinism obtains.

So that is my argument for (4) from the Moral Luck Argument. Premise (3) I argue for elsewhere.[181] Premise (1) is what I take to be a standard definition of Moral Luck, and even if it isn't, it can be accepted as a stipulative definition because it is the definition I appeal to in my defense of premise (2) which is the only remaining premise left to defend. That defense is the next and final step of the presentation of my argument.

2. The Impossibility of Moral Luck

I say that moral luck does not exist. This much the Moral Luck Argument requires. It is common to go further and say that moral luck is impossible. The idea is that it is a conceptual truth that there is no moral luck. Is this correct? I am not sure. Remember that the Moral Luck Argument is not one that I endorse. But, like the claim that determinism entails luck I think that a good defense can be given of the claim that moral luck is impossible. I will present two arguments for the claim that moral luck is impossible which build from a principle I argued for earlier, that praise

[181] Chapter 1

and blame are forms of benefitting and harming (**PBBH**).

To some, such arguments are bound to seem misguided because they constitute an attempt to explain something more fundamental by appeal to something less fundamental. The thought is that it is obvious, or knowable a priori, or deeply intuitive that moral luck does not exist, and that it is unlikely or impossible to find any principles which are more obvious or intuitive than the claim that moral luck does not exist which could be used to argue that moral luck does not exist. I do not agree, for two reasons. The first has to do with the fact that, as we have seen, the concept of luck is extremely complex and that initially appealing accounts of luck fail. Typically, those who assert the obviousness of the claim that Moral Luck does not exist also endorse either an account of luck on which X is a matter of luck for Y if X is outside of Y's control or an account on which X is a matter of luck for Y if X is a matter of chance. As we saw both of these accounts are false. Given that the commonly accepted accounts of luck are false how can we trust the declarations of obviousness or direct intuitiveness of claims involving luck? The second is that I have known many quite clever people who claim that moral luck exists. That is by itself sufficient reason to reject the thought that it obviously doesn't or that intuition tells us it doesn't. So I think an argument must be offered.

2.1 Two Arguments for the Impossibility of Moral Luck

The fundamental reason to reject the possibility of moral luck is that moral luck would be unfair if it existed. This might sound Pollyannaish. Certainly unfair things exist. Why should moral luck be an exception? It has to do with the kind of fact would obtain if moral luck existed. If moral luck existed then people could be legitimately praised or blamed for something that is just a matter of luck. For it to be legitimate to praise or blame someone they have to deserve it, so if moral luck exists then people deserve to be praised and blamed for what is just a matter of luck. It is

because the fact that would obtain if moral luck existed is a fact about the nature morality that it matters whether or not it would be unfair if the fact obtains. I will argue that if people sometimes deserve to be praised or blamed for what is just a matter of luck, then people sometimes deserve to be treated unfairly. What I take to be impossible is for anyone to deserve to be treated unfairly. I am not making the strong claim that it is never right to treat people unfairly. I am open to the possibility that sometimes consequentialist concerns can outweigh concerns of fairness. What I don't think is the consequentialist considerations that would make unfair treatment right could make unfair treatment deserved. Or, more simply, I think that if X deserves to be treated Y-ly then it is fair to treat X Y-ly, and that any moral theory which does not endorse this claim is false.[182] If this is right then the impossibility of moral luck is moral fact, if anything is.

I have two arguments for the claim that praising and blaming people for what is a matter of luck is unfair. Both appeal to **PBBH**, the principle that ff X praises Y, X benefits or does good for Y, and if X blames Y then X harms Y, that I argued for in the last chapter. The first argument is called the **Argument from Justice**, and it says:

> J1. **PBBH**
> J2. X does not deserve to benefit from or be harmed by what is just a matter of luck for X.
> J3. So X does not deserve to be praised or blamed for what is just a matter of luck for X.
> J4. For it to be legitimate for X to be praised or blamed X must deserve to be praised or blamed.
> J5. So it cannot be legitimate for X to be praised or blamed for what is just a matter of luck for X.
> J6. Therefore there is no moral luck.

The second argument is called the **Argument from Fairness** and it says:

[182] This is a material conditional, and so any moral theory which denies that anyone deserves anything is one that does not deny this clam linking fairness and desert. Such views are very likely false, I think, and false because of this denial, but they are not false for the same reasons that moral theories which

F1. Every morally relevant property of a person's actions or will counts as either a benefit or a harm to that person.
F2. If the moral quality of a person's actions or will is just a matter of luck, then that person ought to feel alienated from the moral quality of her actions and will.
F3. If someone ought to feel alienated from a harm then that harm counts as an affliction or a deprivation, not a failure, and if someone ought to feel alienated from a benefit, then that benefit counts as a gift, not an achievement.
F4. **PBBH**
F5. To harm someone because they suffer from an affliction or deprivation and to benefit them because they have received a gift are both unfair.
F6. So it is unfair to praise or blame someone for what is just a matter of luck.
F7. So X does not deserve to be praised or blamed for what is just a matter of luck.
F8. For it to be legitimate for X to be praised or blamed X must deserve to be praised or blamed.
F9. So it cannot be legitimate for X to be praised or blamed for what is just a matter of luck.
F10. Therefore there is no moral luck.

The **Argument from Justice** and the **Argument from Fairness** differ in how they get from **PBBH** to the claim that it is unfair to praise or blame someone for what is just a matter of luck. The **Argument from Justice** makes the move by appeal to a single principle **J2**, which says that X does not deserve to benefit from or be harmed by what is just a matter of luck. The **Argument from Fairness** appeals to **F4,** which says that it is unfair to benefit or harm someone because they have already received a gift or deprivation, as well as several principles about alienation.

Before moving on to discuss the details of these two arguments, I need to say something about fairness and justice. Both concepts, that of justice in particular, have received attention from moral and political philosophers, and so it is incumbent upon me to say what account of fairness and justice I am appealing to in making these arguments. The short answer is that I could appeal to any reasonable account of those

---

deny that desert implies fairness are. No-desert views fail to explain all the moral phenomena, while views which deny that desert implies fairness fundamentally misconceive central moral concepts.

concepts, because the claims I make about justice and fairness ought to be considered part of what any good account of the concepts needs to account for. In other words, any account of fairness or justice according to which premises **J2** and **F5** are false is thereby a bad account.[183]

So these arguments should be consistent with any good theories about what is fair and what is just. Are they? I think so. The conclusions certainly are. Prominent theorists about justice and fairness have assumed that it is unfair for people to suffer from simple bad luck.[184] This may seem to only account for half my conclusions, the half dealing with bad luck rather than good luck, but in the context of distributive justice, which is the context in which justice and fairness are most often discussed, ensuring that people do not suffer from bad luck is going to involve ensuring that other people do not benefit from good luck, at least most of the time.[185] As for how well my two arguments comport with adequate theories of justice I should point out that justice only gets mentioned in the title. The argument itself is about people getting what they deserve or not, and it is only because of my attachment to a view of justice on which justice is just a matter of people getting what they deserve that I name the argument as I do. Here again I think that the claims I make accord with any intuitively correct

[183] There are various reasons this could be so. It could be that **J2** and **F5** are part of, or follow from part of, the common sense concepts of justice and fairness. It might be that these premises are implied by all of the normative theories which are likely to be true, and so that they should be considered part of the concepts of fairness and justice. There are likely other stories besides these to tell about why it is that some claims are part of the data to be explained in philosophy and some claims are controversial claims we use good theories and accounts to prove or disprove. I do not want to take a stand on which story is the best. I wish only to stand on the assumption that there is a good story out there.

[184] See Rawls (1971), Cohen (1989), Dworkin (2000)

[185] To allow people to benefit from good luck would be to place the people who did not have that good luck at a comparative disadvantage. There will of course be some cases in any theory of distributive justice that is not perfectly egalitarian where some benefits from good luck will be allowed. In Rawls' (1971) case it would be after the demands of the second principle of justice are met, while in Dworkin's (2000) case it would be after equality of resources has been established. But the fact that they would allow some uncorrected for benefits from good luck does not change the fact that there is just as much a presumption, based on concerns of fairness and justice, against benefiting from good luck as from suffering from bad luck. It is worth noting that after these basic minimums, however specified, have been met suffering from bad luck will be allowed as well, but that there will have to be a good explanation provided to those who so suffer about why they are allowed to suffer.

account of what people do and do not deserve.

2.2 The Argument from Justice

J2 seems overly strong. Sometimes people do deserve to be benefitted or harmed because of something that is just a matter of luck. In the last chapter I discussed the lottery winner and said that the lottery winner does deserve to be rewarded for correctly guessing the lottery numbers, even though that was just a matter of luck. The lottery winner had a derived desert claim to the winnings, as opposed to a basic desert claim. So **J2** needs to be revised to say that X does not basically deserve to be benefitted or harmed by what is just a matter of luck. **J4** also needs to be revised in light of this revision to **J2**, from claiming that the legitimacy of praise and blame depends on X deserving to be praised and blamed to claiming that the legitimacy of praise and blame depends on X basically deserving to be praised and blamed.

Is it true both that legitimacy of praise and blame involves basic desert and that the desert it involves is basic desert? I think both claims are correct. I want to look first at whether the desert involved in moral responsibility is basic desert and not derived desert.

2.2.1 Defending J4

What would it mean for the desert involved in moral responsibility to be derived desert? One way for this to be true would for constructivism about moral responsibility to be correct. Constructivism is a position in metaethics which states that some or all moral truths are true in virtue of human beings have certain dispositions to agree to regulate their behavior in accordance with those truths, or to behave as if they were true.[186] Constructivist moral theories can differ about how

[186] For obvious reasons constructivist theories of moral truth are most naturally applied to deontological moral theories which make little to no room for axiological claims in the moral theory.

many moral truths are constructed and how many are not. It is not obviously clear that all moral truths could be constructed, at least not while meeting the conditions that constructivist place on moral truths. After all constructivists are all committed to some form of internalism about moral judgments, and it is not clear how all moral reasons could be constructed while still being such as to necessarily produce some motivation to follow them, at least in rational people. To see this imagine a constructivist having to answer the question, 'Why should we care about what principles all rational people would decide to abide by?' The constructivist might say that you have a natural impulse to care about what other people think, but this is implausible. To say that one ought to care what others agree to because this shows them the kind of respect they deserve, is to appeal to a moral principle in order to justify the moral claims of the construction procedure, which suggests that this moral demand for respect is not itself constructed. Another option would be to say that moral norms are grounded on other kinds of norms about which we need not accept motivational internalism, such as logical or epistemological norms.

As long as constructivism is not committed to all moral norms and facts being constructed, I think it remains plausible to treat the desert claims involved in moral responsibility as matters of basic desert. It seems absurd to think that we could, simply by agreeing, make someone morally responsible for something. If Alex robbed the bank, then there is no way that we could, simply by agreeing to it, make it the case that Richard is responsible for robbing the bank, and no one, constructivists included, thinks that we could. If we all decided that people could be responsible for involuntary movements, wouldn't that just amount to our being mistaken about what responsibility is? It seems that the basic principles of moral responsibility are also just basic moral principles. So if some moral principles cannot be constructed, principles about moral responsibility are a likely candidate.

But while I think that constructivism about moral responsibility is likely false, in the end it doesn't matter whether it is true, because if it is true it is going to have to accept some distinction similar to that of basic versus derived desert. Certainly the thoroughgoing constructivist cannot accept precisely that distinction, because there is no such thing as basic desert on her theory. But in order for constructivism to allow for plausible first-order normative theories, it is going to have to accept a distinction which is nearly extensionally equivalent. After all there is a real distinction between desert claims that come about because of some particular person promising something and those that cannot arise in that fashion. There are certain kinds of desert claims which are not up to us in any straightforward way, and there are those which are. The constructivist will think that, in the end, all desert claims are up to us in some way or other, but if they do not make a distinction like the one I have made they will not be able to explain the priority of certain desert claims over others. For example, duties to country cannot outweigh duties to respect basic human rights. Duties that arise from promises cannot outweigh duties to alleviate severe deprivation. What explains this? I think it is that one set of desert claims derive their force from the other set, and so the latter set cannot be overridden by the first. And constructivist can accept this claim about priority or fundamentality, and simply explain what makes the fundamental set of desert claims fundamental in some way other than I do. I say that they are fundamental because they are not up to us. Perhaps the constructivist will say they are fundamental because while they are up to all of us taken together, they cannot be up to any of us taken individually, or in any group smaller than that of all human agents or rational agents. But such an explanation is necessary, and that explanation will have two map on to the common sense distinction between basic and derived desert to a significant extent. So **L2** and **L5** could be revised to talk about the fundamental kind of desert, whatever account of that kind of desert one favors.

2.2.2 Defending J2

The two key premises of the **Argument from Justice** are **J2** and **PBBH**. I have already defended **PBBH**, so all that has to be added to the defense of the **Argument from Justice** is **J2**. So why believe that X does not deserve to benefit from or be harmed by what is just a matter of luck for X? For one thing, this principle seems like common sense. When people have their homes destroyed by a natural disaster or contract a terminal illness, we say, in some contexts, that they do not deserve it.[187] Now there could be many things that we are responding to about cases like natural disaster and disease, but I think what we are responding to is the fact that it was just a matter of luck. To see this, consider the following contrasting case:

> *The Farmer and the Investor:* Manuel, a poor farmer, lives in the shadow of a volcano. He was born there and has never had the money to leave the family farm for somewhere safer. The volcano has not erupted in many years, but it is due for another eruption at any time. Jonas is a foolish, rich young man who has decided to invest all his wealth in the development of a mineral extraction plant in the shadow of the same volcano. The price of investment is low because everyone else is avoiding the project due to safety concerns, while the potential payoff is very high because of the quality of the mineral deposits. Predictably, the volcano erupts, destroying Manuel's home and farm, and the entire mineral operation along with all of Jonas' wealth. Both Jonas and Manuel have lost all their worldly possessions.

I think that Manuel clearly does not deserve to have lost all his worldly possessions, while I think that it might be the case that Jonas did. I think the explanation for this is that losing everything was a matter of bad luck for Manuel and it wasn't for Jonas. Jonas took a serious risk, knowing that more likely than not he would lose everything he had, and his taking that risk was essential to the strategy he had adopted to put his wealth to use. That provides good reason to think it was not a matter of luck that

[187] I think the context in which we are most likely to say and think this is in the case of children who suffer in this way.

Jonas was wiped out, even though there was an element of chance present.[188] What else could explain the differences in desert, if not the differences in luck? Both lost all their worldly possessions because of a natural disaster over which they did not have control. Neither was the intentional victim of suffering.

Because of cases like this I think that **J2** is very plausible as long as we keep in mind that it is basic desert we are discussing. Do I have an explanation of why **J2** is true? No. While I denied that the impossibility of moral luck was a brute moral fact, I do have to appeal to such brute moral facts somewhere, and I think **J2** is a better candidate than the impossibility of moral luck. And while I think that explaining the impossibility of moral luck by reference to **PBBH**, **J2**, and **J5** does amount to giving a good argument, in the Moorean sense, I can understand why someone might think that the impossibility of moral luck is no less obvious than **PBBH**, **J2** and **J5**. It is for this reason that I have included the **Argument from Fairness**.

2.3 The Argument from Fairness

As a reminder let me restate the **Argument from Fairness**:

> F1. Every morally relevant property of a person's actions or will counts as either a benefit or a harm to that person.
> F2. If the moral quality of a person's actions or will is just a matter of luck, then that person ought to feel alienated from the moral quality of her actions and will.
> F3. If someone ought to feel alienated from a harm then that harm counts as an affliction or a deprivation, not a failure, and if someone ought to feel alienated from a benefit, then that benefit counts as a gift, not an achievement.
> F4. **PBBH**
> F5. To harm someone because they suffer from an affliction or deprivation and to benefit them because they have received a gift are

[188] I say only good evidence because I think there are ways of filling in the story so that Jonas' losing everything is a matter of luck. If Jonas invested the money because of some real practical necessity, such as needing to raise enough money to pay a ransom, and the only way to get the return on the investment he needed was to invest in the volcano project, then I think that his losing everything is a matter of bad luck. If, on the other hand, Jonas was merely trying to make enough money to move on to another larger exploitive project, then it is not a matter of luck that he lost everything, and, I take it, he deserved to lose it all.

both unfair
F6. So it is unfair to praise or blame someone for what is just a matter of luck.
F7. So X does not deserve to be praised or blamed for what is just a matter of luck.
F8. For it to be legitimate for X to be praised or blamed X must deserve to be praised or blamed.
F9. So it cannot be legitimate for X to be praised or blamed for what is just a matter of luck.
F10. Therefore there is no moral luck.

F10, **F9**, **F7**,and **F6** all follow from other premises, so they are of no interest in evaluating the argument.[189] **PBBH** I have already defended, and **F8** is the same premise as **J4**, so it has already been defended as well. That leaves **F1**, **F2, F3** and **F5** as the central premises in this argument.

Of those four I think only **F1, F2,** and **F3** need a significant defense. **F5** strikes me as a truism. How could it be fair to harm someone for the reason that they had already been harmed? How could it be fair for someone to benefit someone just because they had already received a benefit or gift?[190] The harms and benefits mentioned in **F5** are unfair because they are entirely gratuitous. But while **F4** is very plausible, **F1**, **F2** and **F3** are very contestable. The concept of alienation figures in both **F2** and **F3** and alienation is almost as slippery a concept as luck.

2.3.1 Alienation

According to my account, to feel alienated from something is to feel that it does not stand in one of several important relations to you, and that it should stand in such a relation to you. After all, to feel that one is alienated from something is to feel that something is amiss. Now it is plausible that for different kinds of objects, there are different kinds of relations such that we are supposed to stand in them to the

[189] On the assumption that the inferences are all valid, which I think they are.
[190] The lottery case is not relevant here though it might appear to be so. In the case of someone winning the lottery we confer a benefit upon them because we promised to do so, not because they already received a gift or benefit.

object, and such that if we do not stand in that relation to the object then we are alienated from the object. Perhaps for activities the relation is that of authorship. Perhaps for talents or capacities the natural relationship is that of controlling them. Whatever the correct details are, what I want to defend is the claim that:

> X is alienated from Y only if X fails to stand in the relation to Y that it is proper or natural for X to stand in.[191]

This gets at one aspect of what we mean when talking about alienation. Another important aspect is the relation the concept of alienation bears to identity.[192] I have identified a necessary condition of alienation, but I am far from having a sufficient condition to give. After all we fail to stand in natural and proper relations to things all the time without being alienated from them. When someone fails to be killed by smallpox they have failed to stand in the normal relation to smallpox, but they are not alienated from smallpox. There are ranges of objects for which it doesn't make sense to talk about our being alienated from them. Some we can't be alienated from because of their insignificance (dying your hair blue doesn't make you alienated from it), and some because it is not clear whether there is a natural or proper relation at all (what is our natural relation to dandelions?), among other reasons. I think that for it to make sense to claim that we are alienated from an object, it has to be the case that the natural or proper relationship we bear to it is one that makes it somehow expressive of our identity. So I will, for the remainder of this paper treat alienation as at least:

> the relation that holds between a person, P, and some object, X, when P does not bear relation R to X, where R is the natural or proper relation

[191] It will likely strike the reader that this account is quite broad. While the account will narrow, it is worth noticing that any adequate account of alienation is going to have to be quite broad in order to accommodate the wide range of things from which a person can be alienated. We speak of people being alienated from elements of their psychology, like addictive desires, from other people, from the products of their labor, from their state, from their religion, etc.

[192] The connection seems pretty well agreed to. Frankfurt started a tradition of talking about desires from which we are alienated as being external desires, and that they have that status because we fail to identify with them.

> for P to bear to X, and where P standing in relation R to X entails that X is expressive of some important aspect of P's identity.

This account is not a lot to go on however, and so I will, for the most part, rely on appeals to examples to justify my claims about alienation. If the reader is not attracted to this loose and very general account of alienation, my hope is that the appeals to examples are sufficiently convincing.

2.3.2 Defending F2: Alienation and Luck

2.3.2.1 The Intuitive Case for F2

Why would it be the case that if the moral quality of Y's actions and/or will is a matter of luck for Y then Y ought to feel alienated from the fact that she her actions and/or will have that moral quality? That would amount to saying that moral quality of her actions and/or will did not bear the proper relationship to Y's identity, because it is a matter of luck for Y.[193] Why would that be the case? Is it because luck is universally alienating? Obviously it is not. It is a matter of luck which parking space I get in the morning, and I am not alienated from my parking space.

The answer is that luck seems to be an all-purpose defeater of the kinds of relations we ought to stand in to evaluatively significant features of our life. In what follows I will present a series of examples to motivate this claim.

The first set of examples are one's in which the circumstances in which a person lives are a matter of luck.

---

[193] On my way of thinking about alienation, it could be that we are alienated from X in one way but not in another. If for some object there are multiple distinct relationships that are proper for us to stand in to it, then we could be alienated from that object with respect to one relationship, but not the other. This fact forces me to make the argument for the No-Moral Luck more complex. Instead of just talking about alienation *simpliciter* I should say that 'if X is a matter of luck for Y then Y ought to be alienated from X at least with respect to one relationship.' This is not an altogether strange way of talking. Suppose that after years of good relationships, two sisters become embroiled in a fight over each other's religious activities. It might be that, qua friends, they are alienated from each other, but that 'she is still my sister.'

> *Working Class Alienation* – Tom works the line at a factory. He works 12 hours a day so that he can earn enough to pay for rent, food and transportation. This work schedule leaves no time for anything other than sleeping, eating, washing his clothes and a few other activities necessary to keep his job. Tom works this way because the factory happened to be hiring when Tom left school. Now that he works there he has little ability to retrain for a different job. He is now stuck in this position. Tom knows people he grew up with whose work situation is quite different. When they left school there were other opportunities available. Tom thinks that if things had gone a little differently for him, if there had been a job opening that would have allowed him to pursue more education while working, that he would be capable of doing much more with his life than simply subsist. In particular Tom is certain he could be a sculptor, greatly desires to be a sculptor and is greatly frustrated by his inability to become one.

It is not enough, for Tom to count as being justifiably alienated from his job and from the material conditions of his life that it is not the life he would choose. Most people, perhaps all people, live a life that is not exactly what they would choose. The feeling of alienation is not the feeling of disappointment. It is important in this example that Tom has some concrete idea about what it is he should be doing. He ought to be a sculptor. That is what he values doing, and he thinks he has the proper skills for it. It is not just any divergence from what one wants out of life that could make for alienation, but this is not just any divergence. To see that imagine the case with a slight change, where what frustrates Tom is how much work he has to do, or how he cannot buy a good mattress, or microwave. These are ways in which his life is less than optimal, by his own lights, but the frustration Tom feels over those divergences would not and could not, assuming Tom is a psychologically healthy adult, constitute a feeling of alienation from his life.

> *Arranged Marriage* – Frieda lives in a culture where marriages are arranged. At a young age it was decided that she would marry Samuel, a member of a prominent family who wanted access to the lands owned by Frieda's parents because of the discovery of oil on, or under, that land. So Frieda is married to Sam because there happened to be oil under her parents' farm. Sam is not a cruel or heartless man. He is simply married to someone he does not love and who does not love

> him. They are polite and respectful to one another, but they do not enjoy each other's company, at least not any more than one would enjoy the company of a casual friend. But, being married, they are required to support one another in their life goals, and share some meaningful goals in life, like raising a family, running a household, planning for retirement, engaging with the community at social gatherings, etc.

What seems clear to me is that if Frieda does not feel alienated from her marriage and the aspects of her overall situation which come along with it, then there is something wrong with the way that Frieda is thinking about her situation. Perhaps she lacks the self-respect to think that core aspects of her life ought to be determined by her decisions. Perhaps she is unreflective and so does not appreciated the fact that different modes of life are possible.

> *Orphan* – Shu is a Chinese orphan, whose parents died in a car wreck. He is eventually adopted by a German couple and goes to live with them in Munich. He is immersed in Bavarian culture. But he could, he thinks to himself from time to time, just as easily have been adopted by a British couple, an Indian couple, or a Chinese couple. This thought makes Shu feel as though the culture in which he is immersed is not really his own, that the activities characteristic of members of that culture that he engages in are not fully wholehearted. It is not that he thinks German culture is somehow not valuable. How could he? Who doesn't like Bach? It is just that he feels as though the culture is not his culture. After all he had parents to whom he was biologically related. He had an ethnic/cultural group in which he would have been immersed, had tragedy not struck.

These are all cases of alienation, and, I would claim, are paradigm cases of it. What I want to highlight about these cases is that in them, it seems pretty clear, the feeling of alienation is caused by the belief that the circumstance from which the person is alienated is just a matter of luck. In Tom's case what was a matter of luck were the material conditions of his life, and this was a matter of luck in virtue of it being a matter of luck which employers were hiring and not hiring when he left school. In Frieda's case that it was a matter of luck that oil was found beneath her family's farm

made it a matter of luck who she was married to. In Shu's case that it was a matter of luck which parents adopted him, and this made it a matter of luck which culture he lived in.

Now imagine that Tom had lost out on the chance of being a sculptor not because of the poor opportunities when he graduated, but because he was lazy the first few years out of school, and did not pursue the opportunities that existed. Would it any longer be appropriate for him to feel alienated from his life, or at least from the fact that he never had the opportunity to get the kind of job that would allow him to develop his ability as a sculptor? I think it would not be. After all it was a decision of Tom's that led to his situation, and for one's life to be shaped by one's decisions is how things are supposed to be.[194]

2.3.2.2 A Problem for F2

So that is the intuitive case for **F2**. I think the intuitive case is strong, but I think that the premise needs to be refined so as to escape an equally intuitive case for denying **F2**. The basic problem is that even bracketing issues about the relationship between determinism and luck, luck seems to pop up quite a bit in ordinary life. And when luck appears, it is not always alienating. In fact there are some very central projects in one's life the success of which depends, for most people, on luck. So, for example, it is for most people a matter of luck that they meet their eventual spouse. Many facts about one's children and the kind of people they will become depend on luck. Whether one is successful in one's chosen career seems to depend on luck. But we are not alienated from our spouses, our children, and our jobs. So **F2** is false, so

[194] Complexities can arise given the nature of the decision. Suppose that instead of years of laziness, Tom got to his current situation because he passed up one great opportunity. When he was presented with the opportunity Tom was not in his right mind, whether because of drinking or fatigue. The decision he made, to pass up the opportunity, was not in accord with his considering beliefs on how to handle such situations, and he cannot imagine what he was thinking when he made the decision. It strikes me as plausible that he should feel alienated from how his life turned out, because in this case the decision did not bear the proper relationship to his identity. In one sense he didn't make the decision at all, it was just the alcohol/weariness talking.

the argument goes.

I think the argument against **F2** fails and I think that to see why one needs to attend to two facts. One is that alienation, according to the definition I have offered, only results when a natural relationship is subverted somehow. The second is that there is a difference between something's being just a matter of luck for a person and something's being partly a matter of luck for that person. In **F2** I mean to be talking about benefits and harms that are entirely a matter of luck. What is more, I think that in the cases under discussion the natural relationship at issue is one that admits of partial luck, just not complete luck.

Lets take a closer look at one of the problem cases for **F2**, the case of one's relationship with a spouse. As we saw earlier with *Arranged Marriage*, having the matter of who will be one's spouse settled just by luck is alienating. What is also almost always true is that it is just a matter of luck whether two people who will become spouses meet. I met my wife in college and my meeting her was almost entirely the result of our happening to pick the same dorm hall to live in. Our meeting was entirely a matter of luck I think. But does that mean our subsequent marriage is entirely a matter of luck, or put another way, is it therefore entirely or completely a matter of luck that we got married? Bracketing concerns about determinism, I don't think so. It was a matter of luck that we met, but from that point on our relationship developed as it did in large part because of choices we made. That the relationship is a product, in large part, of our choices, it seems to me the relationship is probably not a matter of luck, not completely anyway. Certainly the fact that a necessary condition of our getting married, our meeting in the first place, was a matter of luck does not imply that our getting married is just a matter of luck. It is a necessary condition of my admittance to the Cornell graduate program that Cornell University was at some point founded, and it is a matter of luck to me that this happened. But my being

admitted was not just a matter of luck.

Now my meeting my wife is more than just a necessary condition of our getting married, it is an essential part of any properly informative story about our relationship. The same is not true of Cornell's founding and the story of my career as a graduate student. The right thing to say here is that because of the luck involved in meeting my wife, my marrying her is partly, but not completely a matter of luck, and that being partly a matter of luck in no way involves our failing to have the relationship natural between spouses. There is a significant difference between the *Arranged Marriage* case and the case of my wife and me. Our choices helped to guide our relationship in a way that Frieda's did not, and this explains why it is not true that we ought to feel alienated from our marriage and Frieda should.

The case of one's relationship to career success is different. Career success is a matter of luck to a far greater extent than one's deep personal relationships are. Whether one gets a job for which many others are competing is just a matter of luck in many labor markets.[195] Whether one advances in one's profession depends in large part on how well others do relative to you. Whether a person keeps her job depends in large part on market forces which are just a matter of luck for almost every individual in the market.[196] It seems as though success in one's career is not just partly a matter of luck, that it is, or approaches, being completely a matter of luck. It is not just that it is a matter of luck that the process leading to career success of failure started as it did. Every step along the way to career success or failure looks like it is a matter of luck that it occurs as it does. But ought we be alienated from our careers? More importantly, ought we be alienated from them for the reasons listed, i.e. the dependence of our success upon the actions of others and the operation of impersonal

[195] The philosophy job market seems to be the paradigm example of luck in job allocation.
[196] The behavior of market forces being a matter of luck is even clearer than natural disasters being a matter of luck, since market forces are, in the west at least, largely driven by investment activities which are only nominally distinguishable from gambling.

market forces.[197]

I think we should distinguish between two different ways to be successful. One is success at the activity one is asked to do in one's career, and the other is career success. The standards for the first kind of success differ from career to career. Authors need to be able to write good books, accountants need to be able to accurately and honestly calculate debt and holdings and the like, managers needs to be able to promote efficiency among their employees. The standards of success for career success are more general, they are the same standards across careers. I have in mind things like advancement in rank or pay, gaining the respect of one's colleagues, acquiring a stake in decision making, etc.

It is career success, conceived of generally, that seems to be mostly a matter of luck, not success in the activity which constitutes one's career. Whether one is a good author does not depend on what your publisher thinks of you, whether you advance in the publishing world does. And while the natural relationship authors have to writing does plausibly exclude that success being mostly a matter of luck, it is not at all clear that the same is true of career success.[198] So I do not think that the kind of luck we have in how our careers go entails that we ought to feel alienated from our success or failures in our career, because the aspects of our career that ought not be matters of luck usually aren't, and those aspects which are usually matters of luck are such that this is not a problem.

So in some aspects of our lives, luck is not alienating. What **F2** claims is that

[197] It is important to make these distinctions, because there might be reasons having to do with oppression and repressions of creativity that make it the case that we ought to be alienated from the way we earn our living, and these reasons don't seem to be directly relevant to luck.

[198] An interesting question is how much the natural relationship depends on the importance someone ascribes to the object in question. So while many recognize that career success depends on recognition from others and because of that don't treat it as very important, some people still do. For them career success is among the central goals of their lives. When career success is given this kind of importance I think that the natural relationship to career success changes and it becomes luck excluding, because career success is, for this person, at the same time a central life goal, and I think that the natural relationship to our central life goals is luck excluding.

when it comes to the moral quality of our actions and will, luck is alienating. The reason why is that our success or failure in our moral projects is, or ought to be, one of the central preoccupations of our lives, and so the general point I argued for earlier, by appeal to examples, covers that moral success.

2.3.3 Defending F3

Now I want to look at **F3**. Why should one think it is true that if someone ought to feel alienated from a harm then that harm counts as an affliction or a deprivation, not a failure, and if someone ought to feel alienated from a benefit, then that benefit counts as a gift, not an achievement? I think that feeling alienated has a regular, though no necessary, connection with other emotions, like pride, shame and guilt. If one feels alienated from some aspect of one's life, it inhibits emotions of this kind in relation to that aspect of one's life. So, for example, if I feel alienated from my job, then it would be very odd for me to take pride in the success of my corporation. It would be strange to run across someone who wore shirts with the company logo on it, attended corporate sponsored activities, and generally speaking loved for things to go well at work, who also claimed to feel that their job was some alien encumbrance on them. It would be similarly strange to find someone who treated the social mores she was raised with as alien, who felt guilt anytime she failed to live up to them. While we might not disbelieve her expression of alienation because of this expression of guilt, we might ask her to see that what made those mores feel alien should also make her not care so much about what they had to say about her. It would seem to us, or to me at least, that she was failing to sufficiently follow the thought or feeling that the mores were alien through.

The two examples I have mentioned are certainly not sufficient to prove that there is a natural connection between these emotions. That would take some empirical work by psychologists, if they could be convinced that such investigations served a

purpose. But I think that the connection is a commonly accepted one, and that is enough for me to go on with confidence to ask why such a natural connection between feelings of alienation and the absence of feelings of pride or guilt should matter. The reason is that the inhibition of guilt and shame is exactly what is appropriate in the case of gifts and afflictions. If the reason you have a great job is entirely because some friend or family member gave it to you through nepotism, then it is a gift. If you walk around feeling proud of yourself for getting that job, then you are either delusional or ridiculous. If you take ill because an epidemic has stricken your town, then to feel guilty for failing to go in for work or be a good host would be irrational. Gifts and afflictions are not things to feel guilty or proud about. I think this fact explains why it is that there is a natural connection between feelings of alienation and the inhibition of feelings like pride and guilt. It is because feeling alienated from a harm or a benefit makes success and failure seem like things that happen to one that we don't normally take pride in them or feel guilt because of them.

2.3.4 A Replacement for F2 and F3?

As I said, alienation is a tricky concept, and while I think that what I have to say about alienation is correct, I also want to provide independent reason to accept the **Argument from Fairness**. Essentially **F2** and **F3** aim to show that if the moral qualities of our actions and/or will are just matters of luck, then our having good moral qualities is like a gift, and our having bad moral qualities is like an affliction. I think this claim is intuitive on its own merits, and to show that I will appeal to pairs of examples, in one of which there is an accomplishment or failure that is brought about in the normal way, while in the other the same result comes about because of luck.

First to cases of accomplishments and gifts:

> *John's Job*: John is looking for a position with a local business. He works on his resume, lines up his references, puts in a personal call to the Human Resources representative, and lands the interview. In

response he buys himself a new suit, so as to look professional and serious, and practices his interviewing skills. The interview goes well, in large part because of John's practicing for it, and John learns that he has the job within a week of interviewing. His new employers tell him how impressed they are at how he presented himself, his references, and his prior experience.

John's Job with Luck: Now imagine that instead of getting back to John telling him how impressed they were with him, his employers let him know that his getting the job was determined by a random number generator. The company is in the business of creating and selling such generators, and as a PR stunt has decided to fill all their job vacancies using it. It was just a matter of luck that John got the job.

Sarah's Knowledge: Sarah is very intellectually curious about a great many topics. This curiosity has led her to pursue advanced degrees in a variety of areas. She studies diligently and always puts her best effort into her work. As a result she has deep knowledge over a wide range of topics.

Sarah's Knowledge with Luck: Sarah has deep knowledge over a wide variety of topics, but not because of curiosity, diligence and hard work. Sarah is very intelligent, has a photographic memory, and, as she lives in Ithaca, is constantly exposed to people with advanced degrees. Through conversation at parties she has managed to develop deep knowledge in all the areas where her friends are well educated.

The claim here is that one would not be justified in taking equal amounts of pride in the accomplishment for both of the pair of situations, in each case, where what distinguishes the pairs is the extent to which luck plays a role in the accomplishment having happened. The explanation for why you aren't justified in taking equal amounts of pride is that in one of the two cases, it just doesn't look like the accomplishment is actually your accomplishment, or even an accomplishment at all.[199] Something similar seems to go for unsuccessful actions and failure.

[199] I will say that I think a person treating a successful action of hers as something other than an accomplishment is an expression of alienation from the action. That I cannot come up with an independent case for the principle argued for by **F2** and **F3** that doesn't at the same time suggest that alienation is lurking in the background is part of the reason why I think that alienation has some explanatory role to play.

> *Game Show*: George is a big trivia fan. He loves to discover and memorize interesting little tidbits. This being the case he applies to be a contestant on Jeopardy. He makes it on to the show and proceeds to play horribly. He wagers poorly, uses a bad strategy in picking categories and amounts, rings in for questions he has no idea about, and allows these mistakes to flummox him so that he misses questions he knows the answers to. He plays an all around poor game.
>
> *Game Show with Luck*: In this case, instead of playing the game poorly, George plays as well as he can. He makes none of the mistakes mentioned in *Game Show*. He wagers intelligently, picks categories he knows best, rings in only if he is likely to know the answer, and keeps his cool. Unfortunately the day he gets on Jeopardy the categories are ones in which he is not very knowledgeable and in which his opponents are well versed. George does his best, but the combination of the opponents and categories he happened to get leads to his losing.

George could and should say, in the luck case, that his not winning the game was no failing of his. It was because of bad luck in the categories that he lost, and while this result was unwelcome it was not his failure. It doesn't look like it was anyone's failure. There is no reason for him to feel shame at his performance or regret about how he played. The most that seems rational here is a wistfulness about how differently things might have gone. This contrasts with the case where George did not win because of his own poor play. There he should feel regret and self-recrimination. After all in that case, it seems obvious, he blew his opportunity.

> *Novelist*: Duncan is a would-be novelist. He has the outlines of a great story in mind. He knows the characters and he knows the plot. The book is going to turn on a crucial chapter which consists of the description of the internal anguish and struggle of the protagonist. The chapter is a tough one to write and the task of writing it has been keeping Duncan from making much progress on the writing of the book. One day while having iced tea in the sunshine, Duncan suddenly realizes how to write the chapter. He begins to write, but the excitement of finally figuring the chapter out is too much for Duncan to handle while writing. He becomes distracted, calling up friends to tell them that he has finally figured it out, having a few shots of whiskey, and imagining what he will say in his interview with the London Review of Books. In all the excitement he neglects to actually write the chapter. When he goes back to his computer after having distracted himself, he finds that he no longer remembers what to write.

> *Novelist with Luck*: In this case Duncan does not allow himself to be distracted with bragging, celebrating and daydreaming. In this case Duncan's building happens to have a fire drill just after he figures out how to write the chapter. Despite his best attempts to keep in mind the way to write the chapter, by the time he gets back up to his apartment the knowledge is gone.

In the first case, Duncan wasted his opportunity to do a good job writing his book. In the second case, it seems legitimate to describe that opportunity as being stolen from him. Here again, as with the previous examples, we see that it is rational, when a successful or unsuccessful attempt occurs, to not feel the emotions characteristic of such a successful or unsuccessful attempt if it is due to luck.

Now as I said I think the best explanation of what I take to be the common intuitions about these cases is that the normal emotional reactions to success and failure are blocked by the unpleasant feeling of alienation from the results. So I stand by the **Argument from Fairness** as presented.

2.3.5 Defending F1

By way of defending **F1** I want to do away with a misreading of it which makes it seem implausible. It can sound extremely egoistic to say that the only morally relevant things about my actions and will involve my being harmed or benefited. This sounds incredibly self-absorbed, as though affairs that do not work to my harm or benefit cannot be morally relevant for me. After all isn't it in cases where no benefit accrues to me that I show most clearly my altruism? And isn't it in cases where it wouldn't have hurt me at all to be kind that I show myself to be heartless or cruel? I do not wish to endorse such a self-centered position. I think that, bracketing concerns about determinism and luck, whenever someone does something blameworthy, they have suffered a harm, because they have failed to respond correctly to the relevant moral decisions. I take it that any time one fails, then one is harmed,

even if one does not have a desire to succeed. By way of illustration think of a drug addict who has been enrolled in a forced rehabilitation program, where drugs are denied him and obstacles to consuming drugs are put in his way. These obstacles and denials are for his own good and effective enough that it takes significant planning and effort on the drug addict's part to get access to narcotics. Suppose the drug addict gets some drugs, consumes them and so fails to rehab properly. Even though the addict had no desire to kick his habit, the fact that he failed at rehab and the fact that it would have been better for him to rehab, makes it the case that he has suffered a harm.

Chapter 4: Real Selves and Reasons Responsiveness

How could a compatibilist respond to the **Moral Luck Argument**? Not by advancing particular accounts of control and ability, for the argument ought to be compatible with any coherent account of those concepts. What she would have to do is argue against either the claim that Moral Luck is impossible or the claim that determinism implies that every morally relevant property of ours holds of us as a matter of luck. How would she do that? I have already pointed out ways in which taking certain positions in moral theory would allow her to deny that Moral Luck is impossible. One might deny **PBBH**, perhaps by endorsing hedonism about the good for human beings. One might deny that if X does not deserve to be treated Y-ly then it is not fair to treat X Y-ly. One might question the connections I have argued for between alienation and fairness. These would all be ways of showing that the **Moral Luck Argument** is unsound, by attacking the moral presuppositions of the denial of the possibility of moral luck.

What would be the problem with this approach taken by itself? It would have the same problem that other piecemeal approaches to compatibilism have, which I mentioned in the introduction. It would not give us any reason to reject other incompatibilist arguments we have seen or give us any reason to be confident that we could respond to future incompatibilist arguments that might be developed.

I think the proper strategy for compatibilists to pursue is one in which particular responses to incompatibilist arguments are derived from an explanatorily powerful, extensionally adequate account of what makes actions free and such that we can be morally responsible for them. We would not want a collection of ad hoc principles which, taken together, give an extensionally adequate account of freedom and moral responsibility. What we want is an account which explains why free

actions are free in a way that makes sense of why all those actions have the moral importance we ascribe to them. After I present my favored compatibilist account, I will explain how it is of use in responding to standard incompatibilist arguments.

I assume that the starting point for any good compatibilists account of free agency and moral responsibility starts with the claim that freedom is connected in some deeply important way to reason. The belief in such a connection has a long history among both compatibilists and libertarians and it has some claim to being a fundamental shared belief about freedom of the will. Both the compatibilist views I will look at in this chapter are organized around the claim that to be free is for one's actions to be controlled by one's practical reason in the right kind of way.

1. Watson's 'Real Self View'

Watson begins the presentation of his view with a qualified endorsement of the classical compatibilist claim that "a person is free to the extent that he is able to do or get what he wants. To circumscribe a person's freedom is to contract the range of things he is able to do."[200] Classical compatibilists like Hobbes, Locke, Hume and more recently Ayer, Schlick, Hobart and Nowell-Smith all endorsed a very similar principle.[201] That principle I will call the conditional account of freedom:

> (**CAF**) X does *A* freely iff if X desires to *A* then X will *A* and if X does not desire to *A* then X will not *A*.

The idea behind **CAF** is that freedom consists in the dependence of one's actions on one's will, where one's will is here construed as an those desires one has that are relevant to the action in question. Obviously this formulation of **CAF** is insufficient. For one thing a person can have conflicting desires, and **CAF** implies that people who

[200] Watson (1975) pg. 205
[201] Ayer (1954), Schlick (1939), Hobart (1934), Nowell-Smith (1948)

have conflicting desires cannot act freely, and this is not true.[202] A further problem is that the identification of one's will with what one desires is overly restrictive. After all on a very common sense way of thinking about desires, I do not desire to go to the dentist and when I do go it is purely because I realize it is good for me to go to the dentist. **CAF** implies that I do not act freely in this case, and that seems not just wrong but close to backwards. There are obvious refinements to make to **CAF** to get it around these problems.[203] A problem remains, however, which cannot be easily dismissed, and it has to do with internal compulsion.

Internal compulsions, like those experienced by those with obsessive compulsive disorder, are plausibly defeaters of freedom. Actions which are explained by internal compulsions are prima facie un-free. The difficulty internal compulsions present is that they do not prevent someone from getting what she wants, rather they determine, in part, what she wants. **CAF** seems to be silent on issues regarding the source of desires. All that matters on this view is the relationship between desire and action, not desires and the mechanisms which produce them.

This problem is one that Watson is aware of. His awareness of the problem explains the qualified nature of his endorsement of **CAF** or something like it. How does he solve the problem? First by introducing a seemingly unhelpful addition to **CAF**:

> (**CAF**') X does *A* freely iff if X *really* wants to *A* then X will *A* and if X *really* does not want to *A* then X will not *A*.

How does adding the '*really*' help? On its own it doesn't. But what it suggests is the

[202] Suppose I both want ice cream and also do not want ice cream. Then, whether I get ice cream or not I will fail one half of the consequent of CAF.

[203] To deal with the first problem one could revise CAF so that it is dependence on the strongest desire that a person's actions must depend on. To deal with the second one could remove talk of desires entirely and replace it with something like 'explanatory attitude', where this could cover evaluative judgments, desires, intentions, etc. It is clear that Frankfurt, in talking about wants and wanting meant to be talking about this larger class of attitudes.

view that we sometimes don't really want to do the things we desire to do. This in turn suggests a diagnosis of what is going wrong in cases of internal compulsion – the subject of the compulsion is pursuing things she doesn't really want even though she has a desire for them – that allows us to keep interpreting cases of un-free actions as cases where the person's actions do not depend on her will.

So that is how employing the concept of 'really wanting' rather than just 'wanting' can help. But until we have an idea of what is involved in an agent really wanting *A* rather than merely wanting *A* we won't know whether Watson is entitled to make use of the benefit of this distinction. There are two questions to be answered here. The first is what makes it the true that 'X really wants *A*' and the second is what psychological states constitutes really wanting *A*.

Before going on to answer those questions, I should back up and point out that this strategy for defending **CAF** is not unique to Watson. Harry Frankfurt pursued a similar line in his 'Freedom of the Will and the Concept of a Person'[204] paper well before Watson. I have chosen to deal with Watson because I think Watson provides a much clearer and more plausible answer to the second question, about what really wanting consists in. That said, Frankfurt provides a much clearer answer to the first question, about what makes it true that 'X really wants *A*'. Frankfurt's answer is that:

> **Identity Claim**: X really wants *A* iff the psychological state which constitutes the token wanting of *A* is either a psychological state X has essentially, or bears the proper relationship to a psychological state that X has essentially.[205]

It is clear how **Identity Claim** works. If we want to know what makes a

[204] Frankfurt (1971)

[205] I attribute the claim that people must have some psychological states essentially to Frankfurt even though he never explicitly accepts it. I do this because I think it is the most straightforward way to take comments (see Frankfurt(2002) among other places) he has made about, for example, Agamemnon and his decision about whether to kill his daughter or give up the invasion of Troy. Frankfurt says that Agamemnon could not survive such a decision because it would require violating volitional necessities. I take it that this commits Frankfurt to the claim that people have their volitional necessities essentially.

wanting a real wanting rather than an alien or intrusive wanting of the kind had by compulsives, Frankfurt will tell us that the real wanting is either an essential part of who we are, or properly connected to an essential part of who we are. How could something be intrusive or alien if it has its source, literally, in our essential nature? How can we intrude or be alien to ourselves?

Now there is much that is not clear in **Identity Claim**. What is the proper relationship that must obtain between essentially held psychological states and non-essentially held psychological states for non-essentially held psychological states to qualify as real wanting? What is the relationship that must obtain between a token wanting and the psychological state that constitutes it? Are psychological states supposed to be literally essential features of a person, or is it enough that they are, for example, a stable constituent of a set of psychological states that is, under some description, an essential feature of the agent? I will deal with these questions in due course, but before I do I need to present the answer to the second question. Watson's answer is:

Agential Rationalism: If X really wants Y then X believes Y to be good.

So really wanting something involves judging that it is good.[206] Why would this be true?

Watson is not clear on this point. He offers **Agential Rationalism** simply as a suggestion, drawn from Plato's moral psychology, which is of help to compatibilists. The problem is that there are other potential accounts of what really wanting involves. Frankfurt, at times, seems to endorse the claim that really wanting Y amounts to having a desire for it that is endorsed by the highest order desire one has relative to the practical question of whether to Y. Bratman endorses the view that when one really

[206] I am treating 'judging X to be good' to stand in the same relation to 'believes X to be good' as 'wanting' stands to 'wants.'

wants Y one has a general intention or policy or plan for life that gives reason giving force to Y. Watson offers an argument against Frankfurt's view, but none against Bratman's.[207] Worse, his argument against Frankfurt is at important points unclear and where the argument is clear, it has been refuted. I will offer a reconstruction of the unclear parts of the argument against Frankfurt because I believe they point to general reasons to prefer **Agential Rationalism** to the alternatives.

1.1 Watson vs. other Real Self Views

Everyone in this debate agrees that it is not sufficient for X really wanting Y that X desire Y. Uncontroversially, people suffering from manic disorders are not responsible for their behaviors. But manic disorders work by generating desires to perform the manic behavior. Addictions sometimes work the same way. In these cases authors who are part of this debate say that the desires in question are alien, or they are such that the agent is alienated from them. The argument for **Agential Rationalism** cannot be that agents cannot be alienated from evaluative beliefs, though it can seem as if this is what Watson had in mind.[208] After all we can be alienated from our beliefs.

> *Racist Liberal*: Consider the liberal who was raised in a racist environment. She believes in the ideals of racial equality and in the irrationality of discrimination based on race. When asked in a moment of cool reflection by another person of her own race, she will affirm these liberal commitments and also make clear that she understands the evidence in support of those beliefs. However, when she actually interacts with members of another race she tends to think things like 'These are violent people' or 'This situation is dangerous'. Later she always renounces those beliefs, but in the moment of actual interaction she accepts them. Her all things considered judgment is that racism is irrational and wrong, but her upbringing has made her such that when

[207] Bratman's version of the real self view takes generalized intentions, rather than evaluative judgments or higher order volitions, serve as the identity constituting attitudes. See especially Bratman (2007)

[208] Watson (1975) says the question which shows that Frankfurt's view is not satisfactory is 'Can't one be a wanton, so to speak, with respect to one's second order desires and volitions?' 'Wanton' is a term of art in Frankfurt's theory, and it is not clear that Watson is using it correctly here, so I think the best way to take his question is 'Can't one be alienated from one's second order desires and volitions?' The answer is yes, but this is not the question that shows that Frankfurt's view is unsatisfactory, unless it also shows that Watson's view is.

> confronted by racial difference she does not take all things into consideration and reflexively falls back on racist beliefs.

I think that the racist liberal's racist beliefs are alien beliefs. I also think there is a good case to be made that they are alien evaluative beliefs. They certainly have evaluative implications. If one thinks that a person is violent one is going to probably also think that a wide range of activities are not ones you should engage in with them. So **Agential Rationalism** should not rest on the argument that we cannot be alienated from our evaluative beliefs. But looking at alien evaluative beliefs provides some clues as to how one could defend **Agential Rationalism**. There is a clear sense in which the liberal racist has racist beliefs. But there is also something odd about her racist beliefs. After all in moments of cool reflection she repudiates them, not just as immoral, but also as incorrect. It seems as though she believes X but also believes that X is incorrect. This has a paradoxical air that I hope can be dispelled.[209]

What I think is interesting is that it is hard to come up with a case of alienation from beliefs that does not involve the agent thinking that the belief is false. In fact I think it is very plausible that when an agent is alienated from a belief, it is because it conflicts with other beliefs which she takes herself to have good evidence for. A second point I want to make is that it is very plausible that no other psychological state aside from belief goes into the proper account of what makes belief rational.[210] It is not as though, when wondering whether it is rational to believe in other concrete possible worlds my desire for a plentiful ontology will play a role in making it rational

[209] I will gesture at a resolution. Let me introduce a very common distinction between dispositional beliefs and occurrent beliefs. Occurent beliefs are one's that a person is currently entertaining and that the person accepts as true. Dispositional beliefs are beliefs that the person is disposed to entertain and accept as true under certain conditions. I think that we should only treat as paradoxical situations where a person has an active belief (A) and also the active belief that (A) is false. Conflicts between dispositional beliefs and active beliefs or between dispositional beliefs and other dispositional beliefs might make the person theoretically irrational, but they are not paradoxical.

[210] I do not mean to endorse the view that whether a person is rational in believing P is determined entirely by other beliefs she has. All I mean to say is that whatever else enters into whether the person is rational in believing P, what desires and intentions the person has will not, except when P is about whether the person has those desires or intentions.

for me to accept Lewis' position. These two facts about belief set it apart from the other kinds of attitudes, like desires and intentions, as better accounts of what really wanting amounts to. Intentions and desires often are rational or irrational because of the relation they stand in to beliefs. If I desire cocaine, that desire is irrational if I believe that cocaine will kill me. If I intend to climb Mt. Everest, that is irrational if I believe that I cannot do it. Something similar goes for the explanation of the alien status of some desires and beliefs. When I am alienated from a desire, part of how I know that is the desire's insensitivity to the beliefs related to that desire. If I think it is irrational to take cocaine, that it is bad to take cocaine, that there is no good reason to take cocaine, that taking cocaine will just make my heart race in an awful way, but still desire to take cocaine, that is good evidence the desire for cocaine is alien and probably the result of an addiction.

Given **Identity Claim** the attitudes that constitute 'really wanting' are in a significant sense basic. The attitudes which constitute a person's really wanting something make it the case that the person is free and responsible, and make it the case that their actions and some other attitudes are non-alien. It would be odd for them to play this role when what makes the attitude rational and non-alien is some different kind of attitude entirely, one that, by hypothesis, the agent does not have essentially. But this is exactly the situation we seem to be in if desires or intentions are substituted for evaluative beliefs in the account of what real wanting amounts to. We are not in the same situation with regards to belief. Beliefs are not made rational or non-alien by other beliefs.

1.2 A Revision to Watson's view

So according to **Agential Rationalism**, **Identity Claim**, and **CAF'** a person, X, acts freely when X's actions counterfactually depend on those evaluative beliefs which in part constitute X's identity. Before moving on to dealing with criticisms of

this view, I want to take some time to clarify it. It is not required on this view that free actions be caused by evaluative beliefs which in part constitute the agent's identity. If, for example, the action were caused by a desire that action could still be free if the desire agreed in the right way with an evaluative belief which in part constituted X's identity.[211] What is required for a desire to agree with a belief? Let us suppose that the right way to think about desires is as an attitude taken towards propositions, just as beliefs are. If a desire for p agrees with an evaluative belief, it will be because the evaluative belief is a belief that p is good. This much at least is required for agreement between desires and evaluative beliefs.

But it seems as though more is necessary. Consider the following case:

> *Alcoholic Wine Enthusiast*: Jon believes that drinking wine is good. Jon is also an alcoholic. On his birthday Jon goes to a winery and drinks wine all day. The explanation for his drinking is the desires that result from his alcoholism. These desires agree in their content with an evaluative belief Jon had. But had Jon been acting from his evaluative belief, he would not have drunk to excess.

Intuitively Jon is not responsible for getting drunk. After all he got drunk because of a compulsive desire to drink. The fact that this compulsive desire agreed in content with some evaluative belief of his is not sufficient to show that he is responsible for his actions. One might think the problem here is that Jon has no evaluative belief in favor of drunkenness, and that he should be required to have one if he is to be held responsible for getting drunk. More generally one might think that for any state of affairs, S, for X to be responsible for S, S must have been caused either by

[211] Watson's position is that we are responsible for those actions caused either by evaluative beliefs or by desires that do not disagree with our evaluative beliefs. The lack of disagreement does not amount to agreement however. It might be that one's evaluative commitments do not speak to an issue in any way, and it is hard to see how one could count as really wanting what one's real wants don't speak to. I agree with Watson that a non-disagreement requirement yields more intuitively correct attributions of responsibility than would an agreement requirement. I just do not see how a non-disagreement requirement can be argued for as anything other than an ad hoc maneuver to get the right results, given the nature of his view.

an evaluative belief that constitutes in part X's identity and that is in explicitly in favor of S or by a desire that is in favor of S. This formula will not work though. Think about another example very close to *Alcoholic Wine Enthusiast* but differing only in that this example Jon is not an alcoholic, just a wine enthusiast. He drinks to excess, but the desire to drink is not compulsive. Here, I think, Jon is responsible for getting drunk, despite the fact that he neither desired nor saw any value in getting drunk. He is responsible for getting drunk because he is responsible for his drinking, and his drinking is what made him drunk, and he had reason to know that this would happen.

I think the right way to complete the account of agreement is by appeal to the modal facts about the desires and beliefs. In Jon's case his desire for drink had nothing to do with his evaluation of wine tasting, and one way to see this is the counterfactual independence of the one from the other. Whether or not Jon judged wine tasting good, Jon was going to have that desire for drink, and vice versa. Their agreement was just a coincidence, and what Jon's case shows is that coincidental agreement is not enough. On the other hand, to demand that at every possible world where X has the belief X also have the desire seems too strong. If someone saves a child from drowning and we find out it is from a general desire to protect the innocent, we would not withhold praise from him if we found out that had he lacked the desire, his moral judgments about the duties of bystanders would have led him to perform the same action. But of course if we are taking seriously the claim that the relevant evaluative beliefs constitute the agent's identity, then to demand that at any world where the agent has the desire she also have the corresponding belief would be to demand too little. By hypothesis she always has that belief, so this demand amounts to just the demand for agreement in content again.

Presumably the problem with compulsive desires is that they normally render practical reasoning irrelevant. No matter what you think about compulsions, you are

unlikely to resist them. So if the desire is such that if the agent were to believe differently, the desire would not have caused the action it is a desire for, then we would be reasonably sure the desire was not a compulsive desire. We face again here the problem that according to Watson the agent is going to have the relevant evaluative beliefs at all possible worlds, or at least all the nearby ones. So what we should say is that the kind of desire, where kinds are individuated by causal source, is such that it would not cause the agent to do what the desire is a desire for her to do, if the desire were opposed by an evaluative belief to the contrary.

1.3 Criticisms of Watson's View

1.3.1 Wolf's criticism

I will deal first with some criticisms of Watson's view that fail before moving on to criticisms which succeed. Susan Wolf has been a consistent critic of Watson's view and views like it that she labels Real Self Views. Wolf spends a chapter of her book, 'Freedom within Reason', discussing the Real Self view, and comes to the conclusion that Real self view is committed to the claim that "the condition that an agent's actions be attributable to the agent's real self is to serve as not just a necessary but also as a sufficient condition of responsibility," and "that any agent who has a real self is responsible at least for any action that is actually governed by her valuational system. Thus any agent who has a real self is responsible for any wholly unalienated actions, for any actions that the agent would, on reflection and in light of relevant information, unqualifiedly regard as actions that are truly hers."[212]

Is it true that Watson must endorse the claim that any action which can be attributed to the agent is one the agent is responsible for? Here it matters what we mean by 'attributability.' It can be used in such a way that to say that X is attributable to A is just to say that A caused X. If that is all that is meant then of course

[212] Wolf (1990) pg. 36

attributability is not sufficient for moral responsibility. If I turn on a light and in doing so ruin some developing pictures, it may very well be the case that the action was caused by my identity constituting evaluative beliefs, but I am not morally responsible for ruining the pictures if I didn't know the room was being used as a dark room.

Of course it is typical for 'attributability' to take on a more elevated meaning and it is not strange to hear people talk about 'deep attributability' or something's being 'really attributable'. What deep, real or true attributability amounts to is not clear to me, and I suspect its meaning is plastic enough that anything which might be taken to matter to assignments of moral responsibility can be built into it. But then it just becomes a stand in for the more immediately contestable term 'responsibility.' So the burden of defending the Real Self view is just the burden of proving that the real self really is the self. After all if Y is attributable to X iff X is responsible for Y, and if, as **Identity Claim** says, the person just is their real self, then of course an action's being attributable to the real self is necessary and sufficient for the person to be morally responsible for the action.

So the first claim that Wolf makes about the Real Self view is either false or trivially true, depending on what you take the term 'attributability' to mean.[213] The claim that, according to the Real Self view, we are responsible for all wholly unalienated actions is one that I do not accept. Watson might have accepted it, but I don't think he should have, for reasons having to do with mundane actions. After all we are certainly not alienated from mundane actions like putting our left shoe on first in the morning. Nor are such actions plausibly in agreement with an evaluative judgment that constitutes, in part, our identity. On the view I am sketching, we would not be morally responsible for those actions. This might seem counter-intuitive, but it

[213] Of course 'attributability' could mean other things, but the only firm meaning I know of is the meaning on which it means something like 'causation' and the only other mean I can glean from its use is of whatever relationship obtains between actions and agents such that the agents are responsible for actions.

shouldn't. After all mundane actions of this sort have no moral significance. An agent could not merit blame or praise for such actions anyway.[214]

But what about morally significant actions which are caused by desires that do not disagree with an identity-constituting evaluative belief, but do not agree with one either? Intuitively a person is responsible for such actions, but on the view as it has been presented so far, it seems as though she cannot be. This appearance is deceptive though. So far I have only attributed an account of free action to Watson and I have yet to present an account of moral responsibility. The general account of moral responsibility that Watson favors is an expressive account of moral responsibility. Expressive accounts can be contrasted with production accounts of moral responsibility:

> **Expressive account of Responsibility**: X is morally responsible for Y only if Y expresses some morally relevant property of X's.
> **Production account of Responsibility**: X is morally responsible for Y only if X is the cause of Y.

Though never stated explicitly, I think most disputants in the free will debate are working with a production account of moral responsibility. The idea is that a person is only responsible for the causal effects of the decisions they make. While the expressive account does not deny that typically if we are responsible for something then we are the cause of it, it does not restrict moral responsibility to such cases. According to the expressive account, states of affairs must express morally relevant

[214]This is not to deny that there are appropriate ways to talk about moral responsibility on which this claim about mundane actions is false. One might think that being morally responsible for something is just a matter of a certain causal relationship holding between a person and an action or a state of affairs, and not the moral character of that action or state of affairs. With such a notion of moral responsibility X's being morally responsible for Y is not sufficient for it to be warranted to blame or praise X for Y. I don't think much hinges on this distinction. The narrower account of moral responsibility that I favor will require that people meet the conditions of the wider account. I tie moral responsibility closely to praise and blame mostly so that it is easier for contestants in the literature to share a concept of moral responsibility. I take it was can all agree that what we are interested in is at least what makes someone a fitting subject for praise and blame.

properties of the agent for the agent to be responsible for those states of affairs. I take it that for Y to express some fact about X it must be the case that Y provides evidence supporting our believing in that fact. If we know that Y could only have been caused by that fact about X, then Y provides evidence for that fact.

What differentiates the accounts? In the kinds of cases I was talking about above, we have actions caused by desires that neither agree nor disagree with evaluative beliefs that constitute X's identity. The morally relevant features of X do not play any role in the causal explanation of these actions, but they are morally significant actions. If Watson were committed to a productive account of responsibility it is not clear how he could justify the intuitive claim that people are responsible for such actions. But Watson is committed to the expressive account, and on that account there is no problem here. After all the fact that X does not have an evaluative belief that speaks to the issue is itself a morally relevant fact about X. A more detailed case will help here:

> *Amoral Agribusiness Executive*: Pete is an executive at a major agribusiness corporation. His business is involved in an attempt to push indigenous farmers off the land their families have held for generations so that they can be replaced with a huge tract of land that will be farmed by transient farmers. Pete does not think that what he is doing is right or good for the world or anything like that. Nor does he think that what he is doing is wrong. He experiences no shame when confronted with the facts about what his corporation is doing, but neither does he self-righteously defend himself. He simply expresses his boredom with the subject and moves on. He does not have any evaluative beliefs concerning this issue.

Certainly the fact that Pete does not care about the fate of the farmers whose lives he holds in his hand is itself a morally relevant fact about him. On the expressive account of responsibility what he is doing to those farmers expresses his lack of appropriate moral concern. And this is a fact about who Pete really is, about what Pete really wants or values or cares about, so holding Pete responsible for not caring is not

problematic on Watson's view.[215]

Before moving on I want to review where we are on Watson's view. Starting from the stated desire of defending:

> (**CAF**) X does *A* freely iff if X desires to *A* then X will *A* and if X does not desire to *A* then X will not *A*.

we moved to:

> (**CAF**') X does *A* freely iff if X *really* wants to *A* then X will *A* and if X *really* does not want to *A* then X will not *A*.

and in an effort to explicate 'really wanting' we were provided with the following two principles:

> (**Identity Claim**) X really wants *A* iff the psychological state which constitutes the token wanting of *A* is either a psychological state X has essentially, or bears the proper relationship to a psychological state that X has essentially.
>
> (**Agential Rationalism**): If X really wants Y then X believes Y to be good.
>
> So the account of free action Watson presents is as following:
>
> **(The Real Self CAF**) X does *A freely* iff if X has an evaluative judgment that in part constitutes X's identity in favor of *A* then X will do *A* and if X has an evaluative judgment that in part constitutes X's identity in favor of not doing *A*, then X will not do A.[216]

and the account of moral responsibility is this:

> (**The Real Self Expressive Account of Responsibility**): X is morally responsible for *A* iff *A* expresses some fact about the evaluative beliefs of X's that constitute, in part, X's identity.

There, then, is the view. What are Wolf's criticisms? She says:

---

[215] This might not do full justice to the intuition that Pete is responsible for what he does. I think the best that this position can do is justify our blaming him for his lack of moral concern. It is not clear at all that we can blame him for destroying the lives of the farmers, and intuitively he is to blame for this. We can blame Pete in response to his hurting the farmers, and his hurting the farmers provides the context for the judgment of callousness. After all Pete is so callous that the destruction of other people's way of life holds no interest for him. But while this result does, in my eyes, save Watson's view from outright absurdity, it probably does not capture all we want with regards to cases of this kind.

[216] I do not say that X must not do *A* if X lacks an evaluative belief in favor of A. This is to reflect the complexity of the moral psychology that Watson is working with, relative to the original classical compatibilist account. If someone acts on a non-compulsive desire that is neither in agreement nor in

> Nonetheless, we sometimes do question the responsibility of a fully developed agent even when she acts in a way that is clearly attributable to her real self. For we sometimes have reason to question an agent's responsibility for her real self. That is, we may think it is not the agent's fault that she is the person she is – in other words, we may think it is not her fault that she ahs, not just the desires, but also the values she does.[217]

Wolf contends that certain forms of insanity lead us to ask these kinds of questions. By way of examples she gives us the Son of Sam murderer, those under the spell of deep hypnosis (as in the novel 1984) and the victims of severe childhood abuse. The contention is that these people are not responsible for "what they are or what they do."[218] Because of the circumstances by which their real selves were produced, Wolf claims, these people are not responsible for what they do, even though they identify with, and are not alienated from their real self.

Now as we have seen the Real Self view is not committed to the claim that agents are responsible for all unalienated actions, so it is not immediately clear that it would claim the people in the cases mentioned above are responsible for their actions. But I think that given certain ways of filling out the details of such cases, the Real Self view would declare the agents in question responsible for what they do. With at least one case I do not see a reason to be the least bit troubled by this position. The Son of Sam strikes me as someone who deserves to be blamed for what he has done, so I will not deal with that case at all. The other two cases that Wolf mentions do present problems. Both are examples which are taken to support the claim that we must be responsible for our character if we are to be responsible for our actions, or, as Pettit has put it, that responsibility has a recursive character.

The Real Self view denies this claim. Wolf correctly attributes this position to the Real Self view but then overreaches in describing what denying that moral

---

disagreement with any evaluative belief X has, then there is no reason to think the person does not act freely. Or, at the very least, there is no reason to think that the person was not free while acting.

[217] Wolf (1990) Pg. 37

[218] Ibid. Pg. 37

responsibility is recursive amounts to. Wolf takes the Real Self theorist to be committed to a view, expressed by Hobart, which Wolf considers to indicate a shallow conception of responsibility. Hobart says, in response to just these kinds of questions about responsibility for character:

> "He did make his character; no, but he made his acts. Nobody blames him for making such a character, but only for making such acts. And to blame him for that is simply to say that he is a bad act-maker."[219]

Wolf's criticism of this claim is that it conflates the conception of responsibility at issue with the kind of responsibility that human beings uncontroversially possess. She says that the property 'being an act-maker' is something human beings share with objects like tires, and events like earthquakes, and so that any conception of responsibility that consists entirely by 'being an act-maker' cannot be the conception of responsibility we are interested. After all when Kant declared that compatibilists gave us the liberty of a turnspit, it was not meant as a way of saying compatibilists were correct, but rather missing the point.

This argument on Wolf's part strikes me as wrong in three ways. One is that it is not at all clear that earthquakes and tires produce actions. If it is part of the nature of an action to be intentional, purposive, or intelligent then of course these events and objects do not make actions. And it seems to me plausible that actions are to be differentiated from other events by just those properties. Further Hobart said that it is the accusation of being a bad act maker that constitutes the meaning of blaming expressions. Presumably it is also the accusation of being a good act maker that constitutes the meaning of praising expressions. So moral responsibility would, on this view, consist in being the producer of actions that have moral qualities. Now many compatibilists of Hobart's day also held a kind of consequentialism that most

[219] Hobart (1934)

moral theorists today recognize as deeply wrong, and shallow even. I believe Wolf's knowledge of that fact seeped into her evaluation of Hobart's claim. But compatibilists are not committed to ethical views on which all it is to be a good action is to produce good consequences. A compatibilist could adopt the claim that it is being produced by good intentions that makes an action good, or some combination of effects and intentions. On such a theory, even if tires could make actions they certainly couldn't make good or bad actions. So the depth that Wolf is looking for would be provided by the depth of the moral theory attached to the compatibilist view. The third problem with this criticism is that it simply gets wrong what Watson's view is. There is no reason to think Watson must commit himself to Hobart's view. Hobart's view of responsibility is clearly a productive account, and it has none of the moral psychological complexity of Watson's view.

1.3.2 Watson's criticism of Watson

But despite the fact that Watson's view escapes Wolf's criticism, and despite the fact that it marks a significant improvement over classical compatibilism, the view still fails, as Watson has noticed.[220] Several problems stem from the fact that only evaluative beliefs can be the source of free and responsible action, and some of them are sufficient to reject the view, though some are not. With regard to the latter we should talk about Watson's claim that "Notoriously, judging good has no invariable connection with motivation."[221] I agree with Watson that this is true, but I disagree in thinking that its truth is a problem for his view. This is just the denial of a strong form of motivation internalism about ethical judgments. But there is no reason to think that Watson's view requires motivational internalism about ethical judgments. The most that Watson needs, as far as I can tell, is the truth of the claim that ethical judgments

[220] Watson has discussed all of these potential problems for his account of freedom. The most prominent discussion is in Watson (1987) but there is also some discussion in Watson (1977).
[221] Watson (1987)

can sometimes motivate, and this claim strikes me as very compelling.

There are two major problems for Watson. The first is that **Agential Rationalism** is an implausibly restrictive account of what it is to really want something. According to **Agential Rationalism** anytime someone has a desire which conflicts with an evaluative belief that in part constitutes her identity, if that desire causes the action then the agent did not act freely. But this is implausible. We can see this by considering two different kinds of cases, about both of which Watson admits his theory has counterintuitive implications; weakness of will and perversity. A person displays weakness of will when she acts contrary to her all things considered judgment about what is best to do.[222] In such cases one's evaluative judgments do not determine one's actions, and so the agent does not act freely, according to Watson. This result is unintuitive. In the case of weakness of will the agent fails to do something that it certainly seemed like she could have done. That the non-weak willed action was within her power is what makes her will weak, rather than simply overmatched. If we conclude that she did not fail freely, it casts into doubt the whole basis for treating weakness of will as blameworthy.[223] It is open for Watson to say that in cases of weakness of will the agent can still be morally responsible on account

[222] Watson's position is that people are not responsible for weakness of will, so it is clear that he takes his theory to imply that any all things considered judgment about what is best to do will either be an evaluative belief that constitutes the agent's identity or, more likely, is entailed by such beliefs. You might wonder why it is that evaluative beliefs that are neither part of the agent's identity nor are entailed by them cannot have sufficient weight in an agent's practical deliberation to determine what the agent's all things considered judgment is. Is it really plausible that if the agent has one identity constituting evaluative belief and 10 non identity constituting evaluative beliefs that are relevant to the deliberation, that it is the identity constituting belief that will prevail? The only way this could be is if identity constituting beliefs were taken by the agent to have reason giving force that non-identity constituting beliefs do not. But such an attitude would only be theoretically rational if the agent took her identity constituting beliefs to be more accurate than her non-identity constituting beliefs. In some cases this will be because the non-identity constituting beliefs are not thought to be accurate at all. This is how it is with alien beliefs like those the liberal racist has. But most of the time it will simply be a difference of degree in the credence the agent gives to the belief.

[223] Or sometimes praiseworthy, as with the Huckleberry Finn case that Nomy Arpaly (2003) has discussed at length. It is probably clearer in such cases of reverse akrasia that the agent is acting freely. In Huck's case it is his natural loyalty which wins out over his racist upbringing, and this sounds like an example of someone freeing himself from the repressive practices of his community, a paradigm case of free action.

of the weak-willed action. After all the fact that someone's all things considered judgment failed to determine her action suggests that something is lacking in the depth of her conviction, and in certain cases this seems as though it could be blameworthy.

But this response fails to satisfy for two reasons. The first obviously is that it is intuitive that what we are blaming people for when they are weak willed is at least sometimes the moral quality of the action they performed, not for the weakness of their commitment to their ideals. There is a difference in blameworthiness between a student athlete who breaks his diet and eats a piece of chocolate cake and policeman who violates his own principles by stealing money from the evidence room. Certainly they both show weakness of will, but one of them does something worse than the other and the accusation is that Watson's view would not allow us to blame the policeman more than the student athlete. The second problem with this response is precisely the fact that it allows us to hold the agent responsible for weakness of will while also claiming that she did not act freely. If she did not act freely, how is it fair to hold her responsible for weakness of the will? She couldn't do any better, so how can be blame her for being weak-willed? This points to a tension in the account. It is very plausible that an action must be done freely for someone to be responsible for it, but while on the account Watson gives of free action it is sufficient or close to sufficient for moral responsibility, it is not a necessary condition. What Watson needs is an account of how actions caused by desires that run counter to one's evaluative beliefs can be free and such that we are morally responsible for them.

This lesson shows up again when considering what Michael Stocker has called perverse actions.[224] Simply put, perverse actions are actions which are performed because they are bad, or wrong, or vile, or unjust. When people act perversely they are motivated by the thought that what they are doing is not what they should be

[224] Stocker (1979)

doing, and so they are acting from a desire that disagrees with their evaluative beliefs.[225] But, just as with weakness of will, perversity of action is not an excusing condition; it is itself something that a person can be blamed for, so Watson's inability to explain how we can justifiably hold people responsible for perverse actions is a significant problem for his view.

By far the most serious problem, what I will call the fragility of identity, has to do not with **Agential Rationalism** taken by itself but with the combination of that claim and **Identity Claim**. Taken together those claims imply that for any agent there will be some evaluative beliefs that in part constitute her identity. The most natural way to take that is that the property of having such a belief is an essential property of the agent's. So if the person stops having those essentially held beliefs they would go out of existence. I take it is as obvious that this is an implausible view. It suggests the absurd view that sometimes a good argument or some compelling evidence could amount to a murder weapon. To avoid such a conclusion Watson would have to claim that the beliefs in question are incorrigible, or at least such that no one would stop believing them. Logical and mathematical beliefs might have this feature. So might beliefs such as 'there is an external world' or 'there are other minds.' The problem is that these beliefs have no practical import. In fact it is difficult to think of any beliefs that have practical import that are such that no one could, having once come to accept it, come to reject it.[226]

---

[225] One might think of perverse actions as a kind of weakness of will. But there are significant differences. In weakness of will the person in an important sense wants to do the right thing, as they see it. The agent has, after all, gone through the trouble of figuring out what she thinks the right thing to do is. That she does not abide by that decision does not show that the deliberation that went into it was some kind of sham. When a person performs a perverse action they do not do so in spite of their moral, ethical or prudential commitments. One can think of perverse actions as whole-hearted in the way that weak-willed actions are not.

[226] Perhaps no one who had once accepted that 'torturing innocent babies for fun is wrong' could stop believing it. So we could, on Watson's view, hold people responsible for not torturing innocent babies for fun, or for any other actions that was so obviously wrong or obviously right. This is obviously not going to be a very large set of types of action, and so the account of responsibility is not going to be satisfying.

Isn't it relevant here that Watson has clearly rejected **Identity Claim**? Whether or not he once accepted it, he certainly came to reject it, and perhaps I have identified the reason why. So why should we not replace **Identity Claim** with a principle that is less ontologically robust, such as the one Watson actually seems to endorse:

> **Practical Identity Claim**: X really wants *A* iff the psychological state which constitutes the token wanting of *A* is either a psychological state that is part of X's practical identity, or bears the proper relationship to a psychological state that is part of X's practical identity.

What is someone's practical identity? Watson says that "the point of speaking of the 'real self' is not metaphysical, to penetrate to one's ontological center; what is in question is an individuals fundamental evaluative orientation."[227] He also says that one's adopted ends, which are, I take it, the content of the evaluative beliefs Watson has in mind, "express what I'm about."[228]

So what about this weaker claim? I think we can tell from the 'ontological center' comment that we should not think of practical identity as identity sans phrase nor as part of our identity sans phrase. So someone could lose their practical identity, or have it change significantly, without ceasing to be. This takes care of the problem of the fragility of identity. Unfortunately we have all new problems. The essentialist claim made in **Identity Claim** was not gratuitous; it did real work. It put beyond question the claim that agent was really committed to the wants identified as real wants. Can this weaker claim do the same job? It isn't clear, partly because the suggestion is so under-described. But there are some prima facie reasons to think it cannot. What **Identity Claim** did was to give reason why it is that compulsive desires were not a source of responsibility, but other wants were. Can **Practical Identity**

[227] Watson (1996)
[228] Ibid.

Claim do the same?

I don't think so. After all the connection between a person and her practical identity is contingent. No matter what practical identity she has, she could have had another one. So what is to prevent someone's practical identity from being an alien or intrusive element in her psychology? There is a very real sense in which the agent is stuck with her practical identity. Imagine an agent trying to get rid of her practical identity. What would that amount to? She cannot abstract from her practical identity and make judgments about whether she wants to keep it or discard it. Without her practical identity she is incapable of making such judgments, as the material necessary for making evaluative judgments just is her practical identity. On Watson's picture we all might have had very different practical identities, but we are stuck with the kind of person we are, and that is all there is to say. I do not think it is a coincidence that the same article where Watson presents **Practical Identity Claim** he also claims to be only explaining part of moral responsibility, attributability. Once we know that someone's actions are attributable to them, we are justified in making aretaic judgments, Watson thinks. Examples of aretaic judgments are 'He is a coward' or 'She is dishonest'. They are judgments about excellences that people have or lack.

If this is all Watson is arguing for then there is no clear sense in which he is taking part in the traditional debate about free will. No incompatibilist should deny that aretaic judgments are still justified in the face of determinism. Whether someone tends to tell lies for bad reasons is something that does not depend on the whole causal history of her actions, and so there is no prima facie compatibility problem to be resolved here. So this way of weakening **Identity Claim** saves Watson's view from one absurdity, only to lead it into triviality.

I cannot see anyway out of these three problems for Watson, and so now I wish to turn to another account of moral responsibility that, while saddled with its own

problems, can help resolve Watson's difficulties. As an added bonus Watson can help solve the problems that this account, the Reasons-Responsiveness account, is saddled with.

2. Fischer and Ravizza's Reasons-Responsiveness View

Fischer and Ravizza[229] present the following view:

> **Reasons Responsiveness**: X is responsible for Y only if Y proceeds from a moderately reasons responsive mechanism that X has taken responsibility for.

Fischer and Ravizza intended for **Reasons Responsiveness** to serve as a specification of a more general and intuitively plausible principle:

> **Control Principle**: X is responsible for Y only if Y is under X's control.

Y's proceeding from a moderately reasons responsive mechanism that X has taken responsibility for is meant as an account of what it is for X to be in control of Y.[230] I am going to sidestep these issues about control, and focus on **Reasons Responsiveness** on its own merits.

2.1 Reasons Responsive Mechanisms and Asymmetry.

There are two aspects of the view that need to be explained. The first is that of a moderately reasons responsive mechanism, and the second is the notion of taking responsibility. What is a reasons-responsive mechanism? Fischer and Ravizza admit that they have little to say, but what can be said is that it is a psychological mechanism which allows agents, to certain extents, to be receptive and reactive to reasons. A strong reasons responsive mechanism is one that, whenever there is sufficient reason

[229] Fischer and Ravizza (1998)

[230] This is only an account of one kind of control, what Fischer and Ravizza call guidance control. I will not argue about whether this is a good account of guidance control, because I see no reason to think there is any such thing as guidance control. That is, I think the distinction Fischer and Ravizza attempt to draw between regulative control and guidance control fails. I have no intuitive grasp of the distinction and see no reason to posit one. I tend to think there is just one kind of control, that the reasons-responsive account of control that Fischer and Ravizza present fails, but also that **Control Principle** is false.

to do otherwise, would be receptive to that reason and would lead to an action in accordance with that reason. A weakly reasons responsive mechanism is one that would in some possible worlds where there is sufficient reason to do otherwise lead a person to do otherwise. It should be obvious why Fischer and Ravizza don't want to make it a necessary condition of a person being morally responsible for her action that it issue from a strong reasons responsive mechanism. That would mean that unless the agent was such that she would always recognize that reason to do otherwise and act on that reason, she would not be responsible for her actions. This condition would rule out moral responsibility for wrong actions, since in every case of wrong action there is sufficient reason to do otherwise that the psychological mechanism is not sensitive to.[231]

It is harder to see why it is that weak reasons responsiveness is not sufficient for moral responsibility. To show why I will borrow an example from Michael McKenna.[232] Suppose Matilda is at a dance, and loves to dance so much that she would not leave the dance floor for $100 dollars. But if she were offered $1,000 she would leave the dance floor. So the psychological mechanism governing her dancing decisions is weakly reasons responsive. But then suppose that Matilda, while she would leave the dance floor for $1,000 would not leave it for $1,001. The problem is that the pattern of effects of the mechanism does not conform to any reasonable pattern, and so cannot count as reasons-responsive at all.[233] So a moderately reasons

[231] This might not be true. It might be that there are tragic cases in which no matter what a person does, she acts wrongly, perhaps because she violates someone's rights no matter what, but in which one of the actions is the best of the wrong actions.
[232] McKenna (2009)
[233] This is not the way I would put the problem. What Matilda's dispositions show us is that she is not responding properly to monetary incentives at all. It is not that she is responding to the sufficient reason to stop dancing when she takes the $1,000. The sufficient reason to stop dancing in that case is that dancing is not worth $1,000. If she were responding to that reason that would mean that she had an understanding of monetary value and some basic facts about numbers. That understanding would rule out the possibility that she would not stop dancing for $1,001. So I don't think we need recourse to counterfactuals to figure out that something has gone wrong here

responsive mechanism is one that need not be strongly reasons responsive, but which has a pattern of counterfactual effects that corresponds to a rational pattern. More specifically what Fischer and Ravizza claim is that responsible actions proceed from mechanisms which are regularly, i.e. rationally patterned, receptive to reasons, meaning that the agent who has that mechanism is aware of reasons in a regular way, and is weakly reactive to reasons, meaning that the agent sometimes acts in the way the reasons recommend.

Before moving on I should briefly deal with a criticism that dogs Fischer and Ravizza.[234] It has to do with what these reasons responsive mechanisms are. To see whether a reasons responsive mechanism is strong, weak or moderate what we have to do is see whether, in other possible worlds that very same mechanism responds appropriately to reasons. As McKenna puts it, we have to hold the mechanism fixed. But to do this, don't we have to know what the mechanism is? This is something that Fischer and Ravizza do not tell us, and it opens them up to potentially absurd conclusions. Assuming you agree that Matilda is not responsible for her dancing behavior, this refusal to specify what the psychological mechanisms are undermines Fischer and Ravizza's ability to explain why. In the Matilda example I had been assuming that the mechanism was 'dance behavior controller', in which case the very same mechanism is present in the possible world in which Matilda is offered $1,001 as the one in which she is offered $1,000. But what if the mechanism really were the 'dance behavior controller for occasions in which monetary offers of even amounts are made'? Then it is a different mechanism in play when she receives the $1,001 offer, and so we cannot establish that she is not appropriately rational via Fischer and Ravizza's test.

How to deal with this? Well for one thing we could just get rid of talk of

[234] See McKenna (2001), Hurley (1999), Shabo (2005), Ginet (2006)

mechanisms altogether, as some other Reasons Responsiveness theorists have.[235] But we could also just provide a specification of what kinds of mechanisms count when it comes to moral responsibility. Now our relationship to our actions is one of causation, and it is reasonable to suppose that our being morally responsible for our actions involves our causing them in the right way. So the reasons responsive mechanisms ought to be the kind of psychological mechanism that could play a role in causing action. Which one's do this? I don't know, I am just a philosopher. It is not the job of philosophers to provide answers to questions like 'what mechanisms in the brain cause action?'[236] The kinds of mechanisms that can figure into our being morally responsible are the kinds of mechanisms that figure in psychological explanations of behavior. Presumably 'dance behavior controller for occasions in which monetary offers of even amounts are made' is not going to enter into a scientific explanation anytime soon.

Fischer and Ravizza differ from other Reasons Responsiveness theorists in thinking that it is the agent's psychological mechanisms, and not the agent herself, who must be reasons responsive. They adopt this position, they say, because it allows them to avoid a position like the one adopted by Susan Wolf and Dana Nelkin. According to Wolf and Nelkin a person is responsible for what she did when it is the case that she could have acted the right way for the right reasons. What this amounts to is the view that a person is responsible when she is able to be strongly reasons responsive, in Fischer and Ravizza's terms. Taken together with a non-conditional account of ability, Wolf and Nelkin's view implies that if determinism is true then no

[235] Fischer and Ravizza have their reasons to talk about mechanisms, but as I hope to show, there are other ways to achieve their goals.

[236] What if the brain is not the mind? More pertinently what if our agency is not just part of our brain? Then couldn't philosophers be the kind of people to answer this question? Perhaps, though I still think they would be at a disadvantage compared to empirically informed psychologists, who would at least know where the gaps in their materialist theories were, and so would know where to look for non-natural causal influence.

one is responsible for behavior that is either wrong or done for the wrong reasons. How does talking about mechanisms rather than agents help avoid this asymmetrical view? Fischer and Ravizza offer Frankfurt style counter-examples to Wolf's claim that in the absence of alternative possibilities no one can be blamed for their actions. The only way that talking about mechanisms rather than agents makes a difference is that this can be useful, according to Fischer and Ravizza, in explaining why we ought to agree with Frankfurt about the force of his examples against PAP.

But there is a better way to avoid this conclusion. One might disagree with the claim that the kind of reasons responsiveness that is required for moral responsibility is the property of responding to actual reasons. Nomy Arpaly provides us with an account on which a person is praiseworthy for her actions when they are done for the right reasons, and blameworthy when they are done for the wrong one.[237] The gloss on acting for a reason, according Arpaly, is that a person acts for reasons X when it is the case that the person is sensitive to reason X or reliably tracks reason X. This seems very much like a mechanism in the sense that Fischer and Ravizza use the term. But what is different is that it is not required, on Arpaly's view, for a person to sometimes get things right, morally speaking, for them to count as morally responsible. This is not just the difference that Arpaly can accommodate the intuition that someone's complete depravity is not an excuse, but also the claim that, with regards to a particular reason, the fact that a person never reacts properly to it is not an excuse.

I think that Arpaly's version of the Reasons-Responsiveness view gets things intuitively right. So why would one prefer versions where people need to sometimes get things right morally, in order for us to be justified in blaming them for getting it wrong? The reason has to do with normative competence. Susan Wolf in particular

[237] Arpaly (2003)

has accused other compatibilist views of disregarding the fact that a moral agent must be normatively competent to count as a responsible agent. At first glance that normative competence is a necessary condition for moral responsibility seems right. We are committed generally to not blaming people for failures when they are not competent. If I do not know how to do calculus I should not be blamed for failing to satisfactorily complete a calculus problem. And it seems as though something similar is true for moral failings. Children faced with moral dilemmas are not blamed for their failures, for example.

Wolf, Nelkin, and let us assume Fischer and Ravizza[238], seek to meet this demand by demanding that people actually be able to react to reasons properly. I don't think that this is the right way to meet the demand for normative competence. Before I say why, however, let me run through what I take to be some facts about how we normally think about normative competence.

For example consider two gymnasts. One is an Olympic class gymnast, the other is an amateur, trying his hand at gymnastics for the first time. When the amateur tries and fails to perform some basic maneuver on the rings, we don't blame him for that.[239] We will recognize that he is not a good gymnast, but we won't think he failed or ought to feel bad about his performance. He lacks the basic competence to be open to the kinds of evaluation actual gymnasts are open to. If he had brought the maneuvers off it would have been nothing but beginners luck, and we don't praise those who show beginners luck as we do those who display great skill.

[238] Wolf (1990) and Nelkin (2008) accept the view that one any occasion of action, it must have been possible, given the past and the laws of nature, that the person act in accordance with the right reasons. Fischer and Ravizza think that the person's mechanisms must be such that they sometimes respond correctly in similar cases. What Fischer and Ravizza are demanding sounds a great deal like the general capacity to do what is right, and so both can be seen to demand the ability, in some sense, to do what is right.

[239] This is not moral blame of course, but it is blame nonetheless. Suppose that the gymnast finds out that because of his lack of effort there is no way he can compete for the national title, an important goal of his that has no moral significance. It would make perfect sense for his parents, coach or friends to say 'You have no one to blame but yourself that you can't compete.'

So far all is as Nelkin and Wolf would have us believe. Competence is a necessary condition for praiseworthiness and blameworthiness. Now consider the Olympic class gymnast. Suppose this competition is one being held in some small high school gym in the middle of Kansas, and the Olympic gymnast is only competing here so that he can meet some minimum number of required events to stay a professional in good standing. But no one else here is even close to his level, and this causes the Olympic gymnast to get lazy and bored. This laziness and boredom leads to his failing to bring off a maneuver on the rings properly. He knows how to do the maneuver, let's say that he has done it hundreds of times successfully, most recently the day before the competition. But he just doesn't put much effort into it and fails. Intuitively he should be blamed for the failure, but it would be wrong, on the basis of a failure which occurred for these reasons, to impugn his status as an excellent gymnast. We have no reason to downgrade our estimation of his gymnastics abilities because of a failure that comes through laziness.

To draw the parallel between the gymnastics case and the case of moral deliberation, it seems to me that the failure of the excellent gymnast corresponds to a failure in reactivity, and not receptivity. His excellence had to do with his knowledge of how to do the maneuver and possession of the physical skills necessary to pull it off, and his failure did not involve anything going wrong with that knowledge or with those skills. The expert knows what to do, but does nothing with that knowledge. His responses to the situation are what is lacking, not his appreciation of what is to be done. The lesson to draw is not that competence isn't necessary for praiseworthiness or blameworthiness, but that competence consists in have the right kind of awareness of what to do and perhaps general skills necessary to perform the action, and does not concern at all one's tendencies to act on that knowledge, or use those skills, properly.

So should we just lop off the reactivity requirement from moderate reasons

responsiveness and say that people are responsible for their actions when those actions are caused by mechanisms which are regularly reasons-receptive? Reasons receptivity seems to me to capture most of what is attractive about Reasons-Responsiveness, but it fails to take account of one important fact; sometimes a failure to have moral knowledge is itself a moral failure.

> *Bad Son:* Pete is tremendously self-absorbed. This leads to him being inattentive to the feelings of others. Further, his self-absorption means he does not realize that he has caused his parents significant sorrow by not communicating with them. He does not realize that his friends feel neglected because he never considers doing something with them that they like but which he does not. In light of this situation it seems pretty clear that he owes his family and friends apologies, and some thoughtfulness. But of course he does not realize this. This ignorance does not absolve him of guilt when he doesn't apologize and doesn't start being thoughtful. This ignorance compounds the problem. He does not realize that he is doing wrong, because he is too concerned with his own fun to think much about the moral quality of his own actions. It is not as though he doesn't know that it has been years since he has contacted his parents, or that he has never gone out with his friends except to places he likes to go. He simply does not appreciate that these facts are of moral interest.

What cases like this suggest is that we need a further refinement. It is not that receptivity as a whole must work regularly, for that would imply that the selfish son is not responsible for the harm he has caused. He knows the facts that make it the case that he has reason to apologize and be especially thoughtful, but because of his moral failings he fails to see them as reason giving. It does not seem right that Pete should get off the hook for hurting people because he is so insensitive to their feelings that it doesn't occur to him that he is hurting them. This insensitivity to the relationship that normally exists between not communicating with people who care about and those people's feelings being hurt is itself a moral failing of his. So, let me introduce now the quite popular claim that normative facts supervene on non-normative facts. I think that what is required for moral responsibility is not regular receptivity to normative

facts, but receptivity to the non-normative facts upon which normative facts supervene. Pete knows all he needs to know in order to realize that he is hurting the feelings of people to whom he has a special duty of care and respect.

But this revision seems like it might go too far. Sometimes moral ignorance is a moral failing, but sometimes it is not. Let us contrast Pete with someone who was raised in such hellish circumstances, perhaps as a child soldier, that he never actually acquires certain moral concepts. He knows the words 'altruism' and 'beneficence' but has never understood them, in part because of the abuse he has suffered, and the complete absence of expressions of such virtues in his life. He is unable to ever recognize situations where reasons for altruism or beneficence are present, not because of an inability to appreciate the non-normative facts upon which the reasons supervene, but because he has never even acquired the relevant normative concepts. Holding someone like this responsible for what they do strikes me as wrong. So the receptivity a responsible person has to the non-normative facts which ground the reasons for action needs to be stronger than her receptivity to the fact that this grounding or explanatory relation holds, but she has to be able, in some broad sense, to appreciate that the grounding or explanatory relationship holds.[240] So let me suggest the following formula:

> **Moderate Reasons Receptiveness**: X is morally responsible for Y when it is the case that Y issued from a psychological mechanism of X's which is regularly receptive to the non-normative facts upon which ground reasons for Y and weakly receptive to the fact that the reasons for Y are grounded on those non-normative facts.

To be strongly receptive to the non-normative facts in question means always or nearly always being aware of them when they are present. What insures that a

[240] I think that this ability can be accounted for entirely in terms of possession of the requisite moral concepts, but I will not argue for that here. There are several ways a compatibilist could account for this ability in a way that does not require the ability to do otherwise.

person is strongly receptive to the non-normative facts will be in part perceptual capacities. But that is not all. Some non-normative facts are only accessible because of theoretical knowledge we have. Perceptual facts alone do not reveal that an utterance is a case of one person ridiculing the other, for example. The receptivity to the fact of the supervenience relation is going to require even more conceptual apparatus, for it does not follow from X being a case of ridicule that X is also a case of rudeness; perhaps it was joking among friends with acerbic senses of humor. A complex web of information sensitive mental attitudes make it the case that one sees some event as having evaluative significance.

2.2 Taking Responsibility

The second aspect of Fischer and Ravizza's view to be explained is the notion of 'taking responsibility for a mechanism.' There is a problem, which only Fischer and Ravizza, seem to be aware of, that faces all reasons-responsiveness views. Mechanisms of the type that I have been discussing can fail to be the agent's own in just the same way we saw that desires or beliefs could fail to be the agent's own in our discussion of the Real Self View. For instance, those suffering from OCD seem to take completely trivial facts to give them reason to engage in time consuming, potentially harmful activities. Another example would be manipulation. If someone only succeeds in responding correctly to reasons because of a mechanism whereby their decision making processes were taken over by someone else, that person would not deserve praise.

Fischer and Ravizza have tried to give an account of what it is for a mechanism to be one's own, in terms of taking responsibility for mechanisms. To take responsibility for a mechanism is to come to believe that the mechanism is causally efficacious, to come to believe that it would be fair if you were blamed or praised for actions which are the result of that mechanism, and to come to believe both of those

things with some evidence. Now it is odd that X's belief about the fairness of holding X responsible is part of the reason it is actually appropriate to hold X responsible. It suggests that someone who tends to honestly make excuses for herself is inevitably going to succeed, and this doesn't seem right. By way of defense Fischer and Ravizza claim that by not seeing one's self as an active moral agent, one makes it the case that one is not an active moral agent. There is no argument offered for this view, they offer it in order to make the intuitive basis of their view clear. In the absence of argument I must report that I see no reason to think that this is true. Consider the hard incompatibilist who thinks that he is not responsible for anything he does, because no one is responsible for anything they do. Now suppose for the sake of argument that hard incompatibilism is false, and that some people are sometimes responsible for what they do. On Fischer and Ravizza's view it looks as though the hard incompatibilist, if he really means it, is not responsible for his actions, even if his view is incorrect. And this seems incredible, and would be incredible even to the hard incompatibilist.[241]

There is a further problem. What is this evidence that the person must be aware of? The evidence for the causal efficacy of the mechanism is not hard to imagine. That is just the efficacy of one's own practical thought on the subject, and we see evidence of that efficacy all the time. But what about the fairness of blaming and praising the person for actions which are caused by the mechanism? What evidence is there supposed to be of this fact? What makes it fair to blame or praise people for their actions? This last question sounds strikingly similar to the problem of moral responsibility, the problem Fischer and Ravizza are trying to answer.[242] So, it seems, their view is that people are responsible for what they do only when they have

[241] See Mele(2000)
[242] If I am right that fairness based incompatibilism is the best form of incompatibilism, then it is the very same question.

good evidence that they are responsible for what they do. This is hardly helpful.[243]

So Fischer and Ravizza's account of what makes a mechanism one's own fails. They need an in principle way to distinguish mechanisms that are one's own from mechanisms which are the vehicle of internal compulsion. And this exactly the job that I earlier said **Identity Claim** was perfect for. As I said earlier I think that these two views solve each other's problems, and I think it is pretty straightforward how **Identity Claim** can help Fischer and Ravizza; mechanisms are our own when they are partly constitutive of our personal identity. Before moving on to give a more robust account of a view which combines the best aspects of both Watson's view and Fischer and Ravizza's view, I want to deal with an argument, offered by Nomy Arpaly, that a reasons-responsiveness view need have no recourse to controversial claims about personal identity, like **Identity Claim**.[244]

2.3 Could Identity Claim Help?

Arpaly says:

> "Compatibilist accounts of self-control tend to become accounts of one part of the self controlling the other another, and there is nothing wrong with that either. The problem starts when we make our notions of praise-and blameworthiness dependent on a notion of self-control, thus, in effect, postulating that some parts of a person's mind are not 'the agent' and some are, or even without postulating such a neat division of the soul, that some things that the agent does are not 'really her actions' while some are. It is hard to argue as to what makes a particular part of the agent deserving of the privilege of being 'the agent', to find out what kind of self-control is the relevant one, and how to accommodate such questions as 'if my great love is calling me to follow her, but my common sense tells me to stay, which is the real me to which I need to submit'."[245]

Arpaly is certainly right that it is difficult to argue for the position that one part of the

[243] For other worries about this see Russell (2000)
[244] As a reminder **Identity Claim** says that X really wants *A* iff the psychological state which constitutes the token wanting of *A* is either a psychological state X has essentially, or bears the proper relationship to a psychological state that X has essentially.
[245] Arpaly (2006) pg. 20

agent constitutes who the agent really is, while other parts do not. The arguments for and against **Agential Rationalism** show how hard it is to come up with good reasons that support one part of an agent's psychology over another. But where Arpaly is incorrect is in her assumption that the account we give of the Real Self is going to provide answers as to what to do in all or even most cases of practical deliberation. It is only this assumption that could justify bringing up the cases of common sense versus love as a criticism of the Real Self view. My own position is that there is no reason to think that the Real Self view is going, all by itself, to answer questions about what to do. For one thing it might be that one's love and one's common sense are both part of the agent's Real Self. There is no reason to think that every case of internal conflict is a conflict between the agent's true self and alien desires. Sometimes agents are actually conflicted, perhaps because two ends they have reason to think are valuable cannot both be realized, or perhaps because two people to whom they are deeply loyal have become adversaries, or perhaps because they simply have conflicting duties. A second point is that it does not follow from the claims I have made about moral responsibility that a person must take the property 'expressing who I really am' to be a reason that speaks in favor of a course of action. So even if at my core I am conservative and hardheaded, that does not seem to be a reason to not follow my great love. Perhaps there is something wrong with being conservative and hardheaded.

Arpaly goes on to say:

> "To say that a severely manic person is not blameworthy for throwing your expensive vase onto the floor, one does not need to speculate that the vase throwing was not really her action but an action performed by her disorder, or that it was not an action at all. It is enough to point out that all the moral concern in the world short of sainthood would not have prevented the a person who suffers from mania from doing something similar, while for a normal person, it takes a marked lack of moral concern to allow such an action. Cigarettes are, by some measures, as addictive as heroin, but unwilling, severe

> heroin addiction can be a blame reducing condition in context in which a cigarette addiction is not. To explain this, there is no need to try to draw a line between irresistible desires and resistible desires, to say that a tobacco driven action is more yours in some way, or to look at differences in the structures of the two addict's wills. It is enough to point out that for a heavy user, heroin withdrawal amounts to torture, while cigarette withdrawal is only slightly uncomfortable."[246]

This amounts to an argument that **Identity Claim** or other claims like it, are not necessary to make sense of and justify our intuitive judgments about responsibility. I think this is wrong. Putting aside Arpaly's own example for a moment, lets look at the kind of case I have been appealing to, that of someone suffering from OCD. Such people suffer from compulsive desires. We need not say that they are irresistible; rather the point is that they have their source in a mental disorder, and they are insensitive to the agent's other beliefs, desires, intentions, commitments, etc. Suppose our obsessive compulsive misses a meeting he had scheduled because he spent hours turning on and turning off the water faucets in his hotel room. What Arpaly would say is that no amount of moral concern would have been sufficient to make the obsessive compulsive wrench herself out of the loop of compulsive behavior and make it to the meeting. So the fact that he missed the meeting tells us nothing about the moral quality of his will. This all seems right, but it seems to be missing the point that we need an account of why the compulsive desires the agent has do not count against the moral quality of his will. If someone had no compulsion and blew off a meeting so that he could do something trivial, we would normally treat that as expressive of contempt for those he was meeting with. It would show that he did not think they were as important as the faucets in his hotel room. Why is this not the case with the obsessive compulsive? Because they suffer from a disorder, one might say. That is true, but why does that matter? The disorder is just a pattern of behavior and thought

[246] Arpaly (2006) pg. 20

caused in a certain way. Certainly it is one that negatively impacts the obsessive compulsive's proper functioning in society, but so does being rude to people, and we don't let people off the hook for being rude because being rude is a pattern of behavior that hurts the rude person. What Arpaly is doing is helping herself to the claim that the desires of the obsessive compulsive do not count when we are figuring out her moral quality of will, but she is doing so without saying why this is the case. She can choose to simply rely on the intuition that this is so, but when other theorists, both compatibilist and incompatibilist, have an explanation of why the intuition is justified, this is a poor foundation to rest on.[247] **Identity Claim** aims to provide just such an explanation.

3. The Combined View

The Combined View as I am calling it, amounts to a significant revision of both views. First let me point out that **Reasons Receptiveness** is importantly incomplete. What I wanted to do, in arguing for it rather than **Reasons Responsiveness** was argue that the normative competency requirement on moral responsibility implied neither that for a person to be responsible for Y it had to have been possible for the agent to have reacted the right way to reasons nor that the agent must sometimes have responded the right way to similar reasons. But the argument for an understanding of normative competence which did not have those implications would not support the claim that the agent could fail to react to reasons at all. I agree with Arpaly that blameworthiness requires some reaction or other to reasons, just not the right reaction. So let us consider:

Revised Reasons Responsiveness: X is morally responsible for Y when it is

[247] What about the drug case? Well while Arpaly is certainly right that how much moral concern is necessary to conquer the addiction matters, we also need to know why the addiction is something that the person faces as something to be conquered. Why is it that the desires of the addiction are not simply to be attributed to me as would any other desire? What makes the addiction something alien to be overcome? The Real Self view has an answer, and Arpaly does not.

> the case that Y issued from a psychological mechanism of X's which is regularly receptive to the non-normative facts which ground reasons for Y and weakly receptive to the fact that the reasons for Y are grounded by on those non-normative facts, and it is the case that X has some measure of reactivity to those reasons.

I have argued that any reasons responsiveness view, whether it is my **Revised Reasons Responsiveness** view or some other needs an account of what makes the mechanisms in question the agent's. My suggestion is that we employ **Identity Claim** to help the reasons responsiveness theorist out. Because **Identity Claim** is about 'really wanting' we will have to present the following account of 'really wanting', meant as a replacement for **Agential Rationalism**:

> **Agential Reasonability**: X really wants Y only if Y is the object of an attitude which is a reaction to the output of a psychological mechanism of X's which is regularly receptive to the non-normative facts which ground reasons relevant to Y supervene and weakly receptive to the fact that the reasons for Y are grounded by those non-normative facts.

Notice that in moving from **Revised Reasons Responsiveness** to **Agential Reasonability** I have changed 'weakly receptive to the fact that the reasons for Y supervene on those non-normative facts' to 'weakly receptive to the fact that the reasons relevant to Y supervene on those non-normative facts.' The reason to do this is that because I have allowed that any reaction to reasons is compatible with moral responsibility, we shouldn't build into the view that in performing Y the agent was reacting to reasons for Y. Consider the following case:

> *Selfish Architect*: Howard is a reasonably intelligent adult living in New York. He owns and operates a large architecture firm. He is famous in the city and is the first person a developer goes to see when planning a new building. As a result he has more clients than he could ever manage himself. Rather than allowing other firms to take jobs he cannot see to personally, he takes every job that is available to him. To cover the work he hires architecture students right out of college and pays them low wages. Now if Howard passed up on some jobs it would allow some of the people who work for him, for low wages, to start their own firms. The different for Howard would be slight. He would be very wealthy and well respected either way. The difference for his

> workers would be significant though. It would allow them to exercise their creativity in the way that they saw fit. It would allow them to live comfortably, and would grant them a measure of economic independence, all of which they lack now. Howard is aware of all these facts, but his reaction to them is to gobble up more and more of the architecture market.

One would not want to say that there are reasons, moral reasons anyway, in favor of Howard's continuing to gobble up the architecture market in New York. What Howard was responding to was the reasons for sharing the market. His reaction to those reasons was to defy them. Howard is selfish, and that selfishness consists in his reactions to the reasons to be selfless.

Revisions are necessary to **Identity Claim** as well. As it stands **Identity Claim** posits a connection between attitudes of a certain type and the identity of the person who has those attitudes. For the same reasons that it made more sense to analyze real wants in terms of belief, we should say that the psychological mechanisms referred to in **Agential Reasonability** are party constitutive of a person's identity. After all the attitudes which constitute reactions to the outputs of those mechanisms are only rational, when they are rational, because of the mechanisms and how well they track the actual normative and non-normative facts. It would be an odd picture of identity which had it that reactions to reasons-receptive mechanisms were part of our identity, but the reasons-receptive mechanisms were not. I do not mean to say that it is only the reasons-receptive mechanisms that are part of our identity. After all the reasons-receptive mechanisms do not contain any necessarily motivational component, and as we are giving an account of how identity relates to action, it seems we need a motivational element to get us from the reasons receptiveness to action. So the Combined View should say that moderately reasons-receptive mechanisms along with a tendency to react in a certain way to the deliverances of that mechanism are part of an agent's identity. Given that we are talking about mechanisms and the reactions to them, the term 'really wants' doesn't work very well now. In its place let

me suggest 'values.' This is a change that is of course congenial to Watson, who all along thought that it was values that were integral to an agent's identity.

At this point I think the preliminaries are well in hand, and I will present the principles to which the Combined View is committed:

> **Responsibility as Expression of Values**: X is responsible for Y iff Y is expressive of X's real values.
> **Valuing as Responsiveness to Reasons**: X values (or disvalues) Z iff X has a psychological mechanism that is regularly receptive to the non-normative facts which ground reasons relevant to Z supervene and is weakly receptive to the fact that those reasons are grounded by those non-normative facts and X has a tendency to react positively (or negatively) to those reasons.
> **Identity Claim'**: X really values Z iff X values Z and X's valuing Z is either an essential property of X's or is properly related to a value X has essentially.

This is not a complete list. We need other principles to fill in some gaps. In particular we need an account of when it is that actions are expressive of real values. I have been treating 'Y expresses Z' as roughly equivalent to 'Y provides good evidence for Z' and when it is that actions provide good evidence for mental states is mostly an empirical matter.[248] With regards to **Valuing as Responsiveness to Reasons** we need an account of what normative facts supervene on which non-normative facts. We also need to know what psychological mechanisms are strongly receptive to those non-normative facts and to the fact that the supervenience relation holds. But these problems are problems for ethicists and empirically minded epistemologists, respectively. To fill in these gaps goes beyond the scope of this paper.

One question I must answer is about the proper relation mentioned in **Identity Claim'**. I dealt with Watson's account of that relation, but given the significant revisions that have been made to the view, it is not clear that agreement is going to always be the right relation. I think the correct picture here is one that admits of

[248] I say 'mostly' because figuring out how to make the folk psychological notions I am employing here precise enough for psychologists to use them seems like an area where psychologists and philosophers can and should collaborate.

several kinds of relationship. Agreement in content is one, but so is explicit endorsement of the kind Frankfurt thought was necessary. If one values Z essentially and one's valuing Z causes one to value W, this too seems sufficient for it to be true that one really values W. There is no reason that I can see to limit one's self to just one among these relations, and so I will not.

This pluralism shows up again in the range of possible forms reactivity to reasons can take. Suppose a person becomes aware that there is good reason to tell the truth to a friend who is laboring under a painful misconception. Assuming the person values honesty, her reaction to this awareness could take the form of a desire to tell the truth, but it could also take the form of a belief that one's duty is to tell the truth. It might be that the awareness of the reason leads right to the person developing the intention to tell the truth. This pluralism about the psychological states that can constitute a reaction to reasons means that the Combined View does not face the problems with Weakness of the Will or Perversity of Action that Watson's view did. Watson had difficulty with such cases because intuitively we can be responsible for weak-willed or perverse actions, but they are by definition actions taken against one's all things considered judgment about what is best. But on the Combined View evaluative beliefs are not privileged the way they are in Watson's view, so the fact that these are actions which go against an all things considered judgment does not imply that we cannot be responsible for them. If one acts in a weak-willed or perversely because of a desire which is a reaction to the deliverance of a moderately reasons-receptive mechanism, then one would be responsible for that action.

For example, consider someone committed to the all things considered judgment that it was wrong to cheat on her taxes but who had a desire to save money which lead her to cheat on her taxes. According to **Agential Rationalism** the all things considered judgment necessarily represented what the agent really wanted, but

the **Combined View** is committed to no such claim. An attitudes representing what an agent truly wants is a matter of that attitude's relationship to identity constituting psychological mechanisms. I have mentioned two forms that the appropriate relationship between attitude and mechanism might take; it could be a causal relationship or one of endorsement. There is no reason to think that an identity constituting psychological mechanism can only stand in the appropriate causal relationship with beliefs or judgments. In the case under consideration the **Combined View** would just say that if the agent is responsible for cheating on her taxes then it is because she really values money more than she values following the law, and that this fact expressed itself in the desire which motivated her action.

Such a story is not the least bit far-fetched. A fundamentally self-interested person could, through exposure to social and educational pressure, develop the belief that loyalty to the state and respect for the law are important. This would not express what the agent truly cares about. Acceding to outside pressure, even if an agent does it by deceiving herself, does not express what the agent truly values. In such a case what the agent really wants and really cares about would have to come about in some way other than the making of a judgment. It could come about through an emotional reaction, such as disgust at the prospect of doing her taxes, or as an impulse to cheat. But it might feel to the agent as though she was acting weakly. She might say to herself, I know I ought to do my taxes honestly, but I cannot help myself. Whether or not she would do so would depend on whether or not she recognized the desire as expressive of what she truly cares about.[249] The more complete the self-deception the more likely the agent is to feel as though she is behaving weakly in pursuing what it is that she truly cares about.

[249] It is important to realize that the agent herself could be mistaken about what her true values are. This is a reason to reject a view, like Frankfurt's, in which an attitude expressed our identity only if we explicitly endorse it. It sometimes takes hard work to figure out what we truly want, and it is hard to see how such a view could accommodate that fact.

A similar story could be told to explain the phenomenology of perverse actions. I will not take a detailed look at a case of perverse action because the **Combined View** is not committed in the first place to the claim that we always act under the guise of the good. A desire might represent what the agent truly cares about without the agent thinking to herself that the desired object is good. Some desires are just impulses, and we do not, in feeling an impulse to do something, think it is good. I do not mean to say that perverse actions are always caused by brute impulses. This might only rarely be the case. But the observation that some desires are just impulses is enough to show that the **Combined View** is not committed to the claim that we only choose something because it is good. Sometimes when we act responsibly the action is motivated by a desire that does not represent its object as anything other than desirable or pleasant. In some cases an agent might find evil pleasant or desirable in this simplistic way. In other cases an agent might perform a perverse action because of a hatred for her society in general, including the moral code it has adopted.

3.1 The Combined View and the Fragility of Identity

The final issue to consider is whether the Combined view is immune to the 'Fragility of Identity Problem.' The answer to that is that it is not immune. No version of the Real Self View is completely immune to it. Accepting **Identity Claim'** brings along with it an account of identity on which a person's identity is fragile to an unintuitive extent. While the Combined View shares this problem with other views of its kind, it does have the virtue of offering an account of identity on which a person's identity is less fragile than it would be if Watson's view were correct.

What made Watson's view implausible was that it was committed to the claim that there were certain beliefs on which one's identity depended. Of course the Combined View avoids that commitment. Valuing can have evaluative beliefs as a component, but no value has evaluative beliefs or even a tendency to have evaluative

beliefs as an essential part. This is another application of the fact that the reactivity component of valuing can take many forms. If someone stopped having certain evaluative beliefs, perhaps because they became convinced of evaluative anti-realism, this would not necessarily mean that they stopped valuing things. Consider someone who values altruism, but then is presented with the specious argument that no one can be altruistic because the tendency to help others is simply an evolved trait which aims at helping the individual survive. If the person is not able to detect the problems with the argument she might well lose her belief that altruism is noble. But what is not plausible is that such an argument will actually cause her to lose all esteem for altruism, if she really valued It in the first place. She will still have a range of emotional reactions, including admiration, respect, pride (if she knows the altruistic person), shame (if she isn't as altruistic as she could be), etc. She will still, if she ever really valued altruism at all, continue to feel the desire to be altruistic. So for this reason values, as the Combined View conceives of them, are more resilient than any attitude.

This does not take us very far however. We also want to know that the moderately reasons receptive mechanisms are more resilient than evaluative beliefs. This seems very likely to be true. Such mechanisms are going to consist of perceptual faculties, and any conceptual apparatus necessary to interpret such information. The most straightforward way to think of these mechanisms is as functions from the non-normative facts to propositions about what reasons there are in favor of a given course of action. We can individuate these functions by their inputs and outputs, so that any causal process that takes set of facts F as inputs and yields set of reasons R as outputs is the same mechanism as any causal process which takes F as inputs and yields R as outputs. This adds an extra layer of resilience to these mechanisms. The way in which a person gets from her awareness of the non-normative facts to her awareness

of reasons is not an essential property of hers. That she gets from those facts to those reasons is, but this function can be realized by multiple causal processes. By way of illustration, consider a person who believed in God and believed that God had forbidden being disrespectful to one's parents and because of these beliefs took the fact that her parents were present to give reason to be respectful. Suppose this person loses her faith. If she still takes herself to have reason to be respectful of her parents, perhaps because she is thankful to her parents for the sacrifices they made, then this will count as the same reasons-receptive mechanism that was in play when she was respectful of her parents because of her religious faith.

So there are at least two ways in which the Combined View presents a more resilient account of identity than does Watson's view. We are still left with the view that there are some values that a person has which the person literally cannot lose, because if the value ceased to exist so would the person. But of course we all know of cases where people we know have had their values change. So either those values are not essentially held, or the cases we are thinking of are not actually cases of a single person experiencing a change in values at all; they might be cases where one person ceases to be and another takes her place.

This last option sounds absurd, but it seems like the **Combined View** is committed to saying that such things could happen. Strictly speaking it needn't carry that commitment. It is open to the compatibilist to simply insist that anytime a person's values change, it was only values inessential to their identity that changed. The compatibilist might even throw in a cliché like 'Leopards don't change their spots.' A general insistence that all change is relatively superficial would, however, be arbitrary. To back it up the compatibilist would have to say why it is that those values are not essential parts of the person's identity, and that is probably going to require figuring out which values are essential parts of her identity. But once the

compatibilist names those values, she is going to face a dilemma. Is it really impossible for any set of circumstances to come along which would make it the case that the person standing before me would lose one of those values?

Suppose we are talking to Miriam, who is a committed racial supremacist. She reliably sees reason not just to favor members of her own racial group over others, but to oppose the efforts of members of any other racial group to achieve anything of significance. Now if Miriam were, after years of such noxious behavior, to suddenly start fighting for racial equality, it would certainly come as a surprise, but would not strike anyone as reason to think that Miriam has been replaced by some other person. And the **Combined View** does not need to disagree. There are psychological stories to tell about Miriam that do not involve any significant change in her essential values. If it were the case that Miriam was only ever a racial supremacist because she both valued being in a group and strongly identifying that group to the exclusion of others, and was exposed to racist ideologies as a child, then as long as she still valued identifying strongly with a group her identity is not compromised. So if the impetus for her to change her behavior was her having become convinced that racial distinctions are unimportant, and that she is most importantly a Christian and not an Anglo-Saxon, then her commitment to racial equality can be seen as an expression of the same value which led her to adopt racial supremacist views, and it is this stable value that helps constitute her identity.

In the overwhelming majority of cases where someone seems to undergo a fundamental evaluative shift a story like this can probably be told. If some story like this could not be told the evaluative shift would seem rationally inexplicable to us. And the fact that the **Combined View** can accommodate all these cases without being forced to say that the person does not survive the evaluative shift marks a significant improvement over other Real Self Views. In Miriam's case her actions were at one

point motivated by an evaluative judgment that her race was superior to all others, and then her judgments on that issue changed. Watson might have said that while Miriam survives it is her practical identity which has changed, but it is hard to know what to do with that diagnosis. If he had accepted the much more straightforward **Identity Claim** he would have been led to the conclusion that Miriam clearly stopped being the same person.[250]

I think it is uncontroversial that **Combined View** implies that personal identity is more resilient than it would be if any other prominent Real Self View were true. The question is whether it allows for identity to be as resilient as identities usually are. At this point I should introduce some terminological clarity into this discussion. I am speaking of identities as though these are things that a person possesses, and on one way of using the term 'identity' that is true. We speak of ethnic identities, professional identities, assumed identities, and all sorts of other identities that a person possesses non-essentially. These are things that a person has, and they play a role in any complete characterization of who that person is, but they are not essential features of that person. So this is not what I have in mind when I talk about identity. When I talk about an identity I mean to talk about that set of properties which make it the case that some person X at time t is the same person as some person Y at time t+n. I am talking about what makes it the case that some person, Miriam in this example, has the personal identity she has. To add more clarity I am talking about what makes the person X numerically identical to the person Y.[251] What the **Combined View** implies

[250] This is of course a very odd thing to say. Who or what is Miriam such that she could remain Miriam yet be another person? This issue could become very complex, but I think for now I am entitled to say that 'Miriam' is the name for the person residing in a particular body, and that **Identity Claim** is committed to the view that multiple people can occupy one body over the course of the life of that body.

[251] I do not want to talk about qualitative identity in the least. For one thing it is not clear to me how one can make qualitative identity and numerical identity come apart, because the following seems to be a property 'being numerically identical with Miriam.' Even if there were a principled way to exclude such properties it seems to me that qualitative identity is just the numerical identity between sets of property types.

is that it is a necessary condition of some person X being numerically identical to some person Y that some of their values are the same.[252] And this further implies that certain changes of value are such that they destroy personal identity.

So let us consider a case where we can stipulate that an value which is essential to someone's personal identity changes. Consider Wallace, a murderer on death row. Wallace killed several people one night after losing his temper at their having beaten him at billiards. He was unrepentant at trial and after being sentenced to death was openly defiant for years in prison. Then he experiences a religious conversion. To keep things simple let us that Wallace was a rather simple minded person before the conversion and that he valued his own pleasure and convenience more than anything else. This value informed all his actions and was at least an essential feature of his character, and if the **Combined View** is right, also of his identity. Then Wallace is exposed to religious literature that condemns him and his old way of life. Here is where the description gets odd. The person we know as 'Wallace' ceases to exist once the religious conversion occurs because it involves an effective renunciation of Wallace's values. There will still be a person we call 'Wallace', and Wallace's body will still exist, but Wallace the person will no longer exist. And this seems bizarre to many people.[253]

---

[252] As I am treating values as mechanisms, this means that identity requires sameness of mechanisms. There are two ways one might take this. One could require only that X and Y both have mechanisms that can be characterized in the same way, so that if X values honesty then Y need only value honesty. This is a more permissive standard than the one I think is correct. The more restrictive standard requires that the mechanisms be numerically identical. What goes into the mechanisms being numerically identical is going to depend on whether or not theses like dualism, idealism or physicalism are correct. For example, if physicalism is true then the mechanisms which constitute valuing are parts of the brain and for one mechanism to be the same as another it would have to be part of the same brain I take it. If dualism or idealism is correct then of course sameness of mechanism would not require sameness of brain. I don't think one needs to take a stand on this issue to defend the **Combined View** as a view about moral responsibility. This question about sameness of mechanisms does not appear to have much to do with the Fragility of Identity problem. The question about whether to allow the more permissive or more restrictive standard for sharing values is relevant to questions about survival of bodily death I take it.

[253] Though many people speak as though they take this possibility seriously. People speak of being born again, or of not being the same person anymore.

This implication of the **Combined View** is not one that can be avoided. The best someone like me can do is bite the bullet while trying to make the experience seem more palatable. And I think much that seems bizarre can be stripped away from this implication. As I already said Wallace's body survives the change in values just fine. This is one way to explain why it seems odd to say that Wallace the person has ceased to exist. His body, which is the aspect of Wallace with which we are most familiar, survives and this might be what explains our insistence that Wallace survives. The survival of that body might by itself be sufficient to make it legitimate to say 'Wallace survives'. The name 'Wallace' could be taken to refer to a cluster of metaphysically independent entities which are in almost all circumstances found together.[254] The body is one of these things that I think is metaphysically distinct from the person that is of more importance, and that is the subject of experiences.

I think the main reason why people find the claim that some values are essential properties of the person who has them absurd is that they take it that if X is not the same person as Y, then X cannot be the same subject of experiences as Y. They take it for granted that if Wallace before the conversion is a different person than Wallace after the conversion then Wallace after the conversion did not experience all the things that Wallace before the conversion did. To deny that Wallace after the conversion experienced committing those murders is absurd, but the **Combined View** need not say such an absurd thing. The **Combined View** makes claims about what makes X the same person as Y, and that is not necessarily the same thing as what makes X the same subject of experiences as Y.

Why am I entitled to distinguish between personhood and subjecthood in this way? This is certainly not what most other philosophers take to be the case. The

[254] Velleman (2002) has suggested a view of this kind. He thinks that the pronoun 'I' when self-applied, can legitimately be taken to refer to many different things associated with the self for which no unified account can be given.

debate over the nature of personal identity is normally motivated by concerns over whether we can anticipate experiencing anything after death or after going into a Star-Trek style transporter device.[255] The reason I disagree with most philosophers about this is first, that it is clear that the necessary conditions for being a subject are not the same as the necessary conditions for being a person, because some animals like dogs are subjects without being persons.[256] Second, personhood has a role in our practical deliberations that subjecthood doesn't and shouldn't. It is in virtue of someone being a person that we can be held to honor our promises to them, enter into friendships and romantic relationships with them, hold them responsible for what they do, etc. Subjecthood does not seem involved in the reasons why it is appropriate to treat persons this way.[257]

So if the facts which made it the case that Wallace before the conversion was the subject he was are not affected by the conversion then Wallace after the conversion is the very same subject as Wallace before the conversion. If this is compatible, as I think it is, with Wallace before the conversion being a different person than Wallace after the conversion, then much of the oddity of the **Combined View** is gone. To say that someone's personal identity can change is to say that the basis for holding them responsible, for being in personal relationships with them, for

[255] Parfit (1984) is probably the seminal work on personal identity and deals heavily with transporter cases and the question about whether we can anticipate having certain experiences. Most literature on personal identity has followed Parfit on this.

[256] Though less obvious I think that it is possible for a person to persist after they are no longer a subject, as with people who suffer from severe brain trauma and have no higher brain activity. Even if they will not regain consciousness it is still the case that they stand in relationships with people that only make sense if they are still persons.

[257] This might seem false because subjecthood is clearly a necessary condition of it being appropriate, for instance, to enter into a friendship with someone. This is so, I think, because subjecthood, or at least having once been a subject, is plausibly necessary for personhood. How could one actually value something without being aware of it at all, or even being capable of being aware of it? But this connection between subjecthood and personhood does not impact my claim that subjecthood and personhood are distinct in the way that the **Combined View** needs them to be. I am claiming that subjecthood persists despite the termination of personhood, which is not incompatible with subjecthood being a necessary condition of personhood.

any claims they make on the basis of past promises, among other things, has been removed.[258] This claim by the **Combined View** is still controversial, but it is not bizarre.

At this point it might seem like I have so restricted the concept of person that I am advocating a view that it is very much like Watson's own view, on which our values are essential to our practical identity, but not our personal identity. This is not the case. It is true that I have admitted that values are not essential to the identity of subjects, but this does not mean I have adopted a position like Watson's. I have simply made clearer what is at stake in personal identity. Many philosophers, perhaps most, agree that personal identity does not involve bodily identity and I have simply argued that personhood is similarly isolated from subjecthood. What Watson advocates is a position on which values are not integral to personhood at all. A practical identity is something that a person has, whereas personhood is not something had by subjecthood or bodily identity, but is, along with them, something we refer to in referring to ourselves. On Watson's view it is a contingent fact that a person has any of the values she has, and on mine it is not. It is a contingent fact on my view that a given subject is the person she is, and so a contingent fact that a given subject has the values it has. But this is not troubling. Assume for a moment that physicalism is true, and that all facts and phenomena are physical facts and physical phenomena. Then X's being a subject is grounded entirely upon certain features of X's brain, and X having the values is grounded upon some other features of her brain. Is it at all worrisome to suppose that consciousness and all the things that go into practical decision making and the having of characteristic conative and cognitive dispositions are manifested in different modules of the brain? Is it troubling that the one part of

[258] It also might imply that no unified narrative account of the person's life is any longer available. It might also that previous allegiances are no longer binding, such as to a political party or nation, so that to work against those groups would no longer count as a betrayal.

X's brain can change independently of the other?

As long as my claim that subjecthood and personhood are distinct is not paired with a further claim that subjecthood has some kind of priority over personhood, then my version of the Real Self View is not importantly like Watson's. There is a final issue related to the resiliency or fragility of personal identity that I want to discuss and that is whether or not a view like the **Combined View** could make do with mere continuity of values rather than sameness of values. If it could then the **Combined View** would be as good as any other psychological account at protecting the resiliency of personal identity. Instead of saying that for person identity to be preserved that sameness of some set of values must be preserved, the **Combined View** would say that for X at time t to be the same person as Y at time t+n, Y's values must be related in some appropriate way to X's values. What is the appropriate relation? The relation that is typical for psychologically continuity accounts of personal identity is a causal relation such that if X has some value V, and Y has some value W, W is appropriately related to V iff there is some non-deviant causal link between V and W.

The benefits of this proposal for amending the **Combined View** are clear, but I think the costs of the proposal include undermining its ability to account for why certain values are not alien but rather belong to the agent. The benefits of the proposal include both greater resiliency of personal identity and the removal of the necessity of saying that a person is essentially committed to some particular value. To see why it brings greater resiliency think again about Wallace. If Wallace's conversion occurred because of some relatively normal causal process then even though all his values changed, they changed in a way that is rationally explicable, then Wallace is the same person before and after, on this proposal. This would eliminate the odd implications of the the **Combined View** when it comes to personal identity without having to appeal to the distinction I proposed between subjecthood and personhood.

The way in which the proposal secures these goods, however, undermines the **Combined View's** response to incompatibilism. The **Combined View**, sans amendment, says that people are morally responsible for their actions when those actions are explained by the person's essential values. To quote Dewey speaking figuratively, "that conduct is ourselves objectified in actions."[259] For any action which we are responsible we can, when questioned about why we are responsible for it, give a clear answer; the action is the outcome of a decision that was guided by our values, what we care about. And the reason why we are permitted to say that the values are ours is because they constitute us, and their constituting us is not incompatible with determinism. This explanation is not available if we reinterpret the **Combined View** in such a way that it implies a psychological continuity account of personal identity. Suppose X has value V at time t, and performs an action explained by X's having that value at time t. In this case X can still say that the action is explained by V. X can even say that the action is explained by X's character, given the common sense connection between our character and what we value. But at this point the explanation runs out. V is an accidental feature of X's. There is no prima facie reason to think that V is not an alien feature of X's character, especially given the fact that X was determined to have V, was powerless with respect to V, has V just by luck, or whatever else the incompatibilist might say. The most the revised **Combined View** could say is that the action at time t was the causal product of a value that played a role in maintaining X's identity. But the way it did so is not incompatible with its being an alien feature of X's identity.

To see this consider Jenny, who, up until she went to college was very open-minded and adventurous. When she reached college she tried drugs and became addicted, and this addiction became the driving force for all that she did. When the

[259] John Dewey(1891) pgs. 160–61.

addiction caused her to hit bottom she received help in the form of entrance into a very strict rehabilitation program. Because of this program Jenny was able to break free from her addiction. Here we have a relatively familiar, non-deviant causal process which seems to preserve psychological continuity over robust psychological change. Jenny went from open-minded and adventurous to a desperate junky to a very severe, disciplined former addict. The fact that all three traits were part of a chain of psychological traits that proceeded in causally regular fashion does not mean that the addiction was not an alien feature of Jenny's psychology while she was addicted. It was alien despite playing a role in maintaining Jenny's identity.

So the **Combined View** must demand sameness of values, not continuity of values, when it comes to its account of personal identity. The account of personal identity it offers is quite controversial, though it does not endorse absurd claims about subjecthood that it might seem to.

4. How does the Combined View Help Compatibilism More Generally?

Recall the reasons I gave for the necessity of compatibilists giving an account of what makes us morally responsible for an action. The idea is that compatibilists needed a theory to unify their responses to various incompatibilist arguments and to explain why those responses were correct. I think that the **Combined View** can do that. In this section I want to explore some ways that the **Combined View** can be of help to compatibilists.

4.1 The Combined View and the Argument from Alternatives

Probably the most straightforward argument for incompatibilism is the **Argument from Alternatives**:

> A1. If determinism obtains then no one can do otherwise.
> A2. If X is morally responsible for doing Y, then X must have been able to do otherwise.
> A3. So if determinism obtains then no one is morally responsible for their actions.

Compatibilists have denied both **A1** and **A2**. Classical compatibilists often endorsed conditional analyses of ability on which determinism was not incompatible with the ability to do otherwise, denying **A1**. Harry Frankfurt has argued that **A2**, also known as the Principle of Alternative Possibilities (**PAP**), is false.[260] Does the **Combined View** make either argumentative strategy stronger?

4.1.1 The Combined View and the Conditional Account of Ability

I don't think that the classical compatibilist strategy of denying **A1** is helpful, but that a similar strategy would be. Classical compatibilists were interested in claiming that 'X is able to Y' just meant 'If X had wanted to Y, then X would have Y-ed.'[261] And this semantic claim is clearly false. This is not the correct account of what the word 'ability' means, as we can see by the number of competent speakers of English who think that it can be true that 'If X had wanted to Y, then X would have Y-ed' even if it is false that 'X is able to Y'. If the semantic claim is correct, then it would be obvious that those expressions were materially equivalent to anyone who understood what 'ability' means. Presenting their argument as based on the meaning of words weakens the compatibilist argument against **A1** and is unnecessary. Instead of making a claim about the meaning of some expression compatibilists should be content making a claim about the real nature of abilities. The meaning of the word is not what is at issue. What is at issue is whether people actually have the ability to do otherwise, and all you need to justify that claim is some position about when people are actually able to do otherwise, not a claim about what people mean when they say 'ability' or what 'ability' means in some natural language.[262]

---

[260] Frankfurt (1969)

[261] See Ayer (1954), Hobart (1934), Moore (1993), Nowell-Smith (1948), Schlick (1939)

[262] Figuring out what the right answers to these claims about meaning are don't seem to be the job of philosopher's anyway. The former question, about speaker meaning seems to be something the speaker has authority over, and the latter question, about word meaning, seems to be a question for linguists.

So if we respond to **A1** with a conditional account of the nature of ability, rather than a conditional analysis of the word 'ability' is there anything that the **Combined View** could add? Yes, I think so. Traditional conditional accounts of ability took the antecedent of the conditional to be the agent wanting something, or deciding to do something, or intending to do something. When the antecedent of the conditional is just the occurrence of a transitory mental state like a desire, or a decision or an intention, the conditional account of ability fails. The reason it fails is the same reason **CAF** failed, internal compulsion.[263] If someone acts as she does because of an internal compulsion, than she is un-free even though she gets what she wants. Internal compulsions present a different problem for conditional accounts of ability. When people act as they do because of a compulsion, be it internal or external, they cannot do otherwise. But some forms of internal compulsion look like they eliminate the ability to do otherwise without making false the conditional, the holding of which is materially equivalent to the person having the ability to do otherwise, according to the conditional account of ability. Suppose I am severely arachnophobic, and at seeing a spider I immediately run away. The psychological effect of seeing a spider is so great that it immediately produces the desire to run away, and completely blocks the formation of any other attitude having to do with the spider. So, because of an internal compulsion that results from a phobia, I run away and cannot do other than run away. This internal compulsion works such that if I have it I cannot face a spider and develop any attitude other than a desire to run away. So the nearest possible world in which I desire, or decide or intend to do something other than run away is a possible world in which I do not suffer from the phobia. In most of those worlds I will successfully do something besides run away, and so the counterfactual conditional in the conditional account of ability is true. So, according to the conditional account of

[263] See Lehrer (1968), Van Inwagen (1983) and Chisholm (2003)

ability I have the ability to do something other than run away, when my running away is caused by an internal compulsion, just because were I to not suffer from that compulsion I would do something besides run away. So the conditional account of ability not only gives the wrong answer when it comes to some cases of internal compulsion, it compounds the problem by offering a terrible explanation. Essentially it says that internal compulsions do not compromise our ability to do otherwise because if we did not have the internal compulsion we would do otherwise.

The **Combined View** can help remedy this problem. The difficulty in the arachnophobia case was that for the antecedent of the conditional to be satisfied the internal compulsion had to be absent. If, in place of beliefs, desires, intentions or decisions, we revised the conditional account of ability so that the antecedent mentioned only the psychological features that are, according to the **Combined View**, essential properties of the person who has them then it would escape the criticism I just outlined. So in place of the 'If X wants to Y, then X will Y' we have 'If X has a moderately reasons receptive mechanism according to which there is sufficient reason to Y, then X will Y.' It would escape the problem because the victim of the internal compulsion would, in the actual world, satisfy the antecedent of the conditional, but fail to satisfy the consequent. In the case of the arachnophobe, while there was no time for her to decide or come to intend or develop a desire to do something other than run away, that does not mean that stable elements of her psychology, such as reasons receptive mechanisms, stop functioning. She is aware of all the non-normative facts that are necessary to see that there is reason to do something besides run away. She is, let us suppose, aware that those non-normative facts make it the case that there is reason to do something other than run way. The way in which the internal compulsion affects her is to prevent her from reacting appropriately to the reasons of which she is

aware.[264] On this revised conditional account of ability, those suffering from internal compulsions are not able to do otherwise because in the nearest possible world in which they have reason to behave otherwise they do not.[265]

4.1.2 The Combined View and PAP

I said before that the **Combined View** was committed to an expressive account of moral responsibility. The **Quality of Will Account** that I defended earlier is an expressive account of moral responsibility, and is, as far as I can tell the only plausible kind of expressive account.[266] I argued earlier that the **Quality of Will Thesis** did not rule out incompatibilism, but it seems as though it might rule out a kind of incompatibilism, the kind committed to **PAP**. According to the **Quality of Will Thesis** a person is morally responsible for her actions when it is legitimate to make a judgment about the person's quality of will on the basis of her actions. So, according to the **Quality of Will Thesis** responsibility is an actual-sequence concept. In other words what matters to moral responsibility is what actually transpired, not what might have occurred. Does this mean that alternative possibilities play no role in how we ought to evaluate a person's actions and the quality of her will in acting as she did? It certainly matters what alternatives the person took herself to have. If a person steals some bread to feed her family, when she took herself to have only the alternative of allowing them to die, her quality of will is better, morally speaking, than someone who steals to avoid the alternative of spending money on the bread that she would rather

[264] This awareness would presumably consist in a dispositional belief about what reasons there are.
[265] This might look as though it generates a problem. Any time we fail to react appropriately to reasons we will meet the conditions in the antecedent but not the consequent. So this sounds like those who behave irrationally or unreasonably cannot do otherwise. In response I think that the conditional account should be again revised so that the test is whether in the nearest possible world that is not the actual world, in which the antecedent is satisfied, the person satisfies the consequent. It does not fall out of this account that irrational or unreasonable actions are ones for which we have no alternative.
[266] The **Quality of Will Thesis** differs from the definition of an expressive account of moral responsibility only in adding that the moral judgments involved in praising or blaming someone must be fair. Any expressive account which did not contain this prohibition and any that has it is a version of the **Quality of Will Account**.

spend at the tracks. But the alternatives the person thought she had come up in any description of the actual sequence, because they are the objects of beliefs that actually played a role in her deliberation.

So the **Combined View**, because it is committed to the **Quality of Will Account** looks as though it implies that **PAP** is false. This is not strictly speaking true. Even if moral responsibility is an actual sequence notion, so that what explains why people are or are not responsible are facts about what actually happened, **PAP** could be true if the absence of alternative possibilities signified that something had gone wrong with the actual sequence.[267] The absence of alternative possibilities would not explain why people are not morally responsible, but it might imply that they are not. If we want to show that **PAP** is false we have to find counter-examples. Luckily counter-examples to **PAP** are easy to find. Harry Frankfurt has provided a template for counter-examples to **PAP**. The following is a counterexample that fits the template:

> *Joseph and the Scientist:* Joseph's neighbor, Mr. White, is a source of constant conflict for everyone he meets. He is spiteful, cruel and inventive in his malice. Joseph attempts to maintain relations with White characterized by politeness, fairness and forgiveness. White's other neighbor, Marcus is a brilliant neuroscientist, and has had enough both of White's misbehavior and Joseph's refusal to condemn him. Joseph has a lot of power in the homeowners committee and if he wished, he could force White to sell his home and leave. So Marcus tricks Joseph into allowing a chip to be placed in his brain which allows Marcus special kinds of control over Joseph's actions. After the procedure Marcus, who is skilled at rhetoric and home surgery, does his best to convince Joseph that White needs to go. Joseph is moved by Marcus' words and decides to make sure that White is forced to leave the neighborhood. Because Marcus knows of Joseph's forgiving nature he keeps close watch on Joseph to make sure he does not lose his nerve. The device in Joseph's brain allows Marcus to closely monitor his mental states. Marcus has identified a particular neural state that precedes changing one's mind. It also precedes reconsidering one's judgments. It also precedes slight doubts about what one is doing that

[267] See Della Rocca (1998), Pereboom (2001)

> are easily brushed aside. So this neural state is a necessary, but not sufficient condition of changing one's mind. If Joseph exhibits this mental state Marcus is going to use the device to force Joseph to go through with it. But the power of Marcus' reasoning along with the long years of abuse from White are such that Joseph never changes his mind. He has White removed from the neighborhood.

Is Joseph responsible for what he did? I think so. He had White removed because he had been convinced by powerful arguments, that it was the right thing to do. But he had no alternative to doing so, because of Marcus' ability to intervene should it have become possible for Joseph to change his mind.

So this is a relatively straightforward counterexample to **PAP**. There is a significant worry, however.[268] The worry is that there is actually an alternative of a kind available to Joseph, the alternative to realizing the neural state that is a necessary condition for changing his mind, and then having Marcus force him to remove White anyway. There is, in other words, a flicker of freedom for Joseph.[269] What is more, it seems like every Frankfurt style counterexample is going to have a flicker of some kind, at least all the plausible ones.

Does this mean that there are no counterexamples to **PAP**? Strictly speaking yes. Does it mean that Frankfurt style counterexamples do not give reason to reject **PAP**? That depends on issues related to what role alternatives are supposed to play in our moral thinking. I said before that **Combined View** is committed to an actual

[268] There are two major objections to Frankfurt style counterexamples. The one I will mention is the Flicker of Freedom response. One I will not mention is the Dilemma response. The Dilemma response says either the causal history of Joseph's action is indeterministic or it is deterministic. If deterministic then the incompatibilist has no reason to share the intuition, and if indeterministic then the scenario is impossible. Perhaps this response posed problems for some kinds of Frankfurt style counterexamples, but not the one I have given. There is no reason why the story I told in *Joseph and the Scientist* is incompatible with indeterminism because indeterminism is not incompatible with there being necessary conditions of action. See Pereboom (2003)

[269] The name 'Flicker of Freedom' is a bit misleading. The response was named by John Martin Fischer(1995), a compatibilist about moral responsibility and determinism who is content to be an incompatibilist about determinism and freedom when freedom is conceived of as the ability to do otherwise. As most would not be willing to abandon the traditional connection between freedom and responsibility, taking the title 'Flicker of Freedom' literally is not fair to compatibilists.

sequence account of moral responsibility, so that the explanation of why someone is morally responsible cannot be that things might have gone differently than they did in some way. I said that this meant that **PAP** could still be true, but only if we treat the absence of alternatives as a sign that something has gone wrong with the actual sequence such that the agent is not morally responsible. So according to the **Combined View**, **PAP** does not serve to explain anything about the nature of moral responsibility. Frankfurt style counterexamples provide evidence for this claim about **PAP.** If we were to assume that the presence of alternatives is what explains why Joseph is morally responsible that would require us to accept that Joseph is morally responsible for removing White from the neighborhood because he might have been forced to remove White from the neighborhood. This is not plausible.

But what if we abandon the explanatory role for **PAP**?[270] Then we are not forced to make absurd sounding claims about what explains why Joseph is morally responsible, while still denying that he would have been responsible had he lacked any alternatives. On this approach the absence of the alternative is infallible evidence that a person is not responsible, without being the fact that makes the person not responsible. This is a perfectly acceptable move to make, but for it to function as satisfactory defense of **A2** in the **Argument from Alternatives** we need to know what fact actually does make it the case that the person is not responsible. What is it about the actual sequence that the absence of alternatives is infallible evidence for?

At this point the incompatibilist is going to have to supply a well worked out actual sequence account of moral responsibility on which moral responsibility is incompatible with determinism.[271] Once she has done this she will have defended **PAP** as far as it can be defended. How is the compatibilist to respond here? Well the

[270] This strategy is popular among source incompatibilists. The first person to commit this response to Frankfurt to print was Della Rocca (1998)
[271] For an example see Robert Kane (1998)

compatibilists certainly can criticize the incompatibilist account of moral responsibility, whatever it is, but this is almost bound to be unsatisfying on its own. Criticize the incompatibilist account and a new revised version will rush to take its place. For the compatibilist to feel secure in her rejection of **PAP** the compatibilist needs her own actual sequence account of moral responsibility to appeal to, one on which the absence of alternatives is compatible with an agent being morally responsible. If such an account is correct then no defense of **PAP** is possible. Essentially, to establish that **PAP** is true we have to be able to derive it from some actual sequence incompatibilist account of moral responsibility, and so the debate over **PAP** should take place on the level of providing accounts like the **Combined View** or comparable incompatibilist accounts.[272]

4.2 The Combined View and Transfer Arguments

Transfer arguments are arguments which employ transfer principles like the transfer of powerlessness principle or the transfer of non-responsibility principle. While there are as many transfer arguments as there are possible transfer principles, I will only talk about two transfer arguments here, the **Consequence Argument** and the **Direct Argument**.[273] The **Consequence Argument** says:

C1. $\Box(P_0 \ \& \ L \rightarrow P)$	(Determinism)
C2. $\Box(P_0 \rightarrow (L \rightarrow P))$	(From C1)
C3. $N(P_0 \rightarrow (L \rightarrow P))$	(From C2 by Rule ά)
C4. NP_0	(Fixity of the Past)
C5. $N(L \rightarrow P)$	(From C3 and C4 by Rule β)
C6. NL	(Fixity of the Laws)
C7. NP	(From C5 and C6 by Rule β)

The 'N' operator here is such that Np means 'no one has or ever had any choice about

[272] It is not the case that only **Combined View** can give compatibilists what they need to show that **PAP** is false. Any actual sequence compatibilist view could. I do think, for reasons I presented earlier, that **Combined View** is the best of the lot however.

[273] Ginet (1966) and Van Inwagen (1980) respectively, were the first to offer these arguments.

p'. Essentially it is a powerlessness operator. To say Np is to say that everyone is powerless with respect to p. Rule ά is an inference rule licensing the inference from '□p' to 'Np'. Rule β is an inference rule licensing the inference from 'Np & N(p→q)' to 'Nq'. 'P_0' stands for 'the state of the world in the deep past'. 'L' stands for 'the laws of nature.' 'P' stands for 'the state of the world at present.' **C1** is just an uncontroversial statement of determinism, and is beyond reproach. While compatibilists have questioned the Fixity of the Past premise(**C4**), the Fixity of the Laws premise(**C6**) and Rule ά, I have no interest in doing so. They all seem obvious to me. What is not obvious is Rule β. Let us call Rule β the **Transfer of Powerlessness** principle.

On the face of it **Transfer of Powerlessness** looks right. If someone loses one's savings because the stock market crashed on account of the bursting of the housing bubble, she can show that she was powerless with respect to the loss of the savings by showing that she was powerless with respect to the housing bubble bursting, and powerless with respect to the fact that the housing bubble bursting made the stock market crash. **Transfer of Powerlessness** models that argument very well. The question is not whether we are ever permitted to employ **Transfer of Powerlessness** in making arguments, but whether we are always allowed to, no matter what kind of events or facts we are talking about. Many compatibilists have denied that **Transfer of Powerlessness** is a universally valid inference principle, which it would have to be for the **Consequence Argument** to be a valid argument.

A similar dialectic presents itself when we look at the **Direct Argument**, which says:

D1. NP_0	(Fixity of the Past)
D2. NL	(Fixity of the Laws)
D3. $N(P_0 \& L)$	(From D1 and D2)
D4. $N((P_0 \& L) \rightarrow P)$	(Fixity of Determinism)
D5. NP	(From D3 and D4 by **Transfer of Non-Responsibility**)

In the context of the **Direct Argument** the 'N' operator is such that 'Np' stands for 'no one is or ever has been morally responsible for p.' 'P_0', 'L', and 'P' mean the same thing in this argument that they did in the **Consequence Argument**. So **D1** and **D2** say, respectively, that no one is or ever has been morally responsible for the state of the world in the deep past and that no one is or ever has been morally responsible for the laws of nature. **D4** says that no one is or ever has been morally responsible for the fact that determinism obtains. Those three premises, along with the inference to **D3**, I will treat as uncontroversial, though again some compatibilists would object. What interests me here is, as with the **Consequence Argument** the transfer principle. The **Transfer of Non-Responsibility** principle says that $(Np \& N(p \rightarrow q)) \rightarrow Nq$. It is formally identical to the **Transfer of Powerlessness** principle, and is in fact implied by that principle along with the assumption that to be morally responsible for something one must not be powerless with respect to it.

What reasons could compatibilists give to reject both of these transfer principles? The question becomes difficult to answer when we realize that we employ something like these transfer principles all the time in thinking about moral responsibility. I already pointed out a case where we make inferences that seem to be instances of the employment of **Transfer of Powerlessness**. We also seem to commit ourselves to the **Transfer of Non-Responsibility** principle when we think about responsibility for consequences. If we want to know whether a doctor is morally responsible for the death of a patient, and that doctor can show that she was not morally responsible for the action which led to the death, because she had been

misinformed by staff at the hospital let us say, and that she was not responsible for that fact that this action led to death, because that is a basic biological fact let us say, then we take the doctor to have shown that she was not morally responsible for the death. We seem to employ the **Transfer of Non-Responsibility** principle when it comes to responsibility for actions as well. If we want to know whether someone is responsible for knocking over a lamp and she can show that she was not responsible for the event which caused the action, because it was a muscle twitch, and that she is not responsible for the fact that the muscle twitch led her to knock over the lamp, then we know that she is not responsible for knocking over the lamp.

What all compatibilists have to deny is that the two transfer principles are valid when applied to events in the deep past, or even events in the near past which occurred before the agent considered which action to perform. How can compatibilists justify that denial when it looks like they appeal to the transfer principles when thinking about responsibility after the moment of decision? I think that here again it is necessary to appeal to an account of what makes people morally responsible against which we can test transfer principles and from which we can derive competing principles that model the everyday inferences that compatibilists and incompatibilists alike are committed to.

So, according to the **Combined View**, are **Transfer of Powerlessness** or **Transfer of Non-Responsibility** valid inference principles? No, they are not. The **Combined View** gives us an account of what makes an action one for which we are morally responsible, and none of those conditions imply contain a commitment to these two principles. According to the **Combined View**, what makes an agent responsible for an action is just a certain relationship existing between an action and essential features of her psychology. Because neither failing to be responsible for the state of the world in the deep past, nor being powerless with respect to it, implies that

we never stand in this expressive relationship to our actions, the **Combined View** implies that both transfer principles are false. What is more the **Combined View** provides alternative principles which can be used to explain why it is that we make inferences that look like we are employing the two transfer principles.

Why don't we consider people morally responsible for consequences of events for which they were not responsible? It is not because non-responsibility transfers along lines of causal or logical sufficiency, but rather because consequences of events for which we are not responsible do not express anything about the moral quality of our will. Why aren't they? Because they lack the proper connection to our will. These consequences are caused, by hypothesis, by events which do not express anything about our will, i.e. cannot be evidence in favor of the judgment that our will has one moral quality rather than another. So how could something caused by an event which expressed nothing about our quality of will come to express something about it? Something similar goes for powerlessness and its role in our moral thinking.

4.3 The Moral Luck Argument and the Combined View

What, finally, does the **Combined View** have to say about the **Moral Luck Argument**? Here I do not have much to say. The previous incompatibilist arguments are all well known, and have been subject to years of compatibilist criticism. All I did in criticizing them was to present, very briefly some of what I take the best criticisms to be and show how those criticisms can be made stronger by tying them to an account of what makes people morally responsible for their actions. There is no established body of compatibilist literature about the **Moral Luck Argument**, because it is, in this form anyway, a new argument.

All I want to do in this section is explore what routes someone with the **Combined View** might take in arguing against the **Moral Luck Argument**. I will by no means show that the argument is unsound, anymore than I showed that about the

more established incompatibilist argument. All I want to do is make it plausible that the **Combined View** has the theoretical resources to respond to the **Moral Luck Argument**, just as all I wanted to show was that the **Combined View** had the theoretical resources to explain why it was that some popular compatibilist replies to traditional incompatibilist arguments work.

So how does one attack my **Moral Luck Argument**? As I said in presenting it, there are two ways to go about it. One is to claim that moral luck is possible, and the other is to claim that determinism does not entail that every morally relevant feature of our actions and will are just matters of luck. I think it is this second claim that compatibilists should make, and it is certainly the claim that the defender of the **Combined View** should make. The reason has as much to do with the inability of the **Combined View** to give reason to reject the claim that moral luck is impossible as it does with the ability of the **Combined View** to show that determinism does not make all the morally relevant features of our actions a matter of luck.

The **Combined View** takes no general positions about the nature of fairness or desert. As such it seems to be in a poor position to provide reason to reject the key premises in either the **Argument from Fairness** or the **Argument from Desert** for the claim that moral luck is impossible. This is not to say that the **Combined View** is committed to those arguments being sound. If the defender of the **Combined View** were forced to accept that determinism entailed that all the morally relevant features of our actions were a matter of luck, then, given that the **Combined View** implies that it can be fair to blame and praise people if determinism is true, she would have to deny one of the claims about fairness and desert made in the **Argument from Fairness** and the **Argument from Desert**. But it is hard to see how she could provide an argument for that on the basis of the commitments of the **Combined View**.

So if the **Combined View** is going to be of help to the compatibilist here, it

will have to be by showing that determinism does not entail that all the morally relevant features of our actions are just matters of luck. How to go about that? The **Combined View** does not mention luck any more than it mentions fairness or desert. As I said I am not going to give anything but a brief case against the **Moral Luck Argument**, so my comments here might be unsatisfying, but the **Combined View**, when conjoined with two plausible principles about luck, shows that some morally relevant features of our actions are not matters of luck, even if determinism is true.

4.3.1 Two Luck Principles

The two principles I have in mind are the following:

> **Luck as Accidental:** *Y can be lucky in having the property X, only if X is not an essential property of Y's*[274]
> **Transfer of Lack of Luck:** *If Y's having some property, Z, is explained by Y's having some property, X, and X does not hold of Y by luck, then Z does not hold of Y by luck.*

It should be pretty obvious how these two principles interact with the **Combined View** to show that not every morally relevant property of our actions is a matter of luck. **Luck as Accidental** and the **Combined View** together imply that every person has some psychological features that are not a matter of luck, those psychological features which constitute the reasons-receptive mechanisms which are essential properties of the person who has them. The relationship that holds between those mechanisms and our actions when we are responsible, according to the **Combined View**, is one of expression. As I have said at many points, the expression relation can

[274] Neil Levy (online manuscript) has criticized this principle, saying that "essentialism about the traits and dispositions at issue is false. Even *if* identity is fixed, across all possible worlds, by the genome, such that any person who did not result from the union of sperm and egg which features in my causal history would not have been me, it is false that there is any psychological trait that I had to have. Psychological traits are not fixed by genes. Indeed, it is not even true that all across all environments, genes or gene-complexes even have a tendency towards particular traits." Very simply I am not offering an account on which the genome is the only essential property a person has. I am not offering a theory on which the genome is an essential property at all. I am offering a theory on which the essential properties are those properties which, in accounts of personal identity, make a person the person they are.

be, and maybe always is, constituted by causal relations. So action A expresses quality of will Q when it is the case that Q caused A. But then Q explains A, and so the relation called for by the **Combined View** to guarantee moral responsibility is just the same relation as the one that guarantees that the action is not just a matter of luck.

The real question is whether these two luck principles are true. And here the **Combined View** does not provide any help. These principles are, if true, conceptual truths about luck, and since the **Combined View** does not contain an account of luck, it is not going to tell us whether the principles are correct. But as long as they are plausible, then the **Combined View** has done a great deal in the defense of compatibilism. Without an account of what makes people morally responsible which, like the **Combined View**, made claims about the essential properties a person has and how they are related to the moral qualities of our actions, the two luck principles would be impotent. When conjoined with the **Combined View** they undermine the **Moral Luck Argument**.

But are these two luck principles plausible? I cannot say much here. What I will say is that **Luck as Accidental** strikes me as intuitively obvious. How could we be lucky to have those properties which make us who we are? Doesn't something, to count as luck, have to be something that might not have befallen you? How can it be luck that you have a certain property when you would not be who you are without it? If the luck is bad, who suffered it? If the luck is good, who benefited? It doesn't seem like it could be the agent. Here again I shall appeal to an example:

> *Patrick and Popular Music*: Patrick was born in 1980. He was too young to appreciate New Wave. He was lucky enough to miss the hair metal phase of popular music, but was not listening to the radio much in the early 1990's when the alternative music movement put very good rock and roll on the air regularly. He started listening to popular music, like most people, in high school when it became socially important to know what was going on in pop music. Unfortunately by the time he entered high school alternative music was already becoming boring and

> repetitive. So Patrick was stuck listening to whiny faux-punk music and bubble gum pop stars.

The question here is not whether it is a matter of bad luck for Patrick that he was stuck listening to pop music on the radio. I am quite certain he was. But there is another question that the example brings up, at least to me, and that is whether Patrick was unlucky to be born when he was. It is a reasonable question to ask. If Patrick had been born a few years earlier he would have caught alternative rock at its height. If he had been born a few years later he would have been in high school just as the New York garage-punk bands replaced pop stars and emo-punk on the airwaves. If Patrick caught himself saying, 'It was just bad luck that I was born in 1980' could he have been correct?

Let us assume, along with the origin essentialists that it is an essential property of a person that they were conceived from the particular sperm and egg that they were in fact conceived from. And so, given some elementary facts about biology, it is and essential property of each person that she was born around the time she was in fact born. If origin essentialism is correct then it was impossible for Patrick to have been born three to four years earlier or later than he was.[275] Someone could have been born to his parents and been named 'Patrick', but it would not have been the same person. And this fact seems obviously relevant to evaluating whether it would have been good luck for Patrick to have been born a few years earlier or later than he in fact was. There is no such thing as Patrick being born in 1977 or 1983, so how could it be bad luck for him that this impossible state of affairs did not come to pass? The absurdity

[275] I said earlier in Chapter Three that I do not accept origin essentialism when it comes to figuring out the essential properties of persons. But on almost any plausible account of what is essential to a person it is going to turn out that she could not have been born at some time distant from her actual date of birth. She would have had very different experiences and a different set of genetic dispositions. It is extremely unlikely on a psychological account of identity, for example, that if Patrick had been born in 1977 he would have been the same person that he turned out to be. There is no reason to think his psychological traits would be at all similar.

becomes more evident the farther forward or back one goes from the date of a person's actual birth. Was it bad luck for everyone in the early middle ages that there were not born in the 20th century, so as to take advantage of the polio vaccine? This is the thought behind **Luck as Accidental**.

To see the intuitive case for the **Transfer of Lack of Luck** principle, imagine that we were wondering whether an athlete's abilities were a matter of luck or not. Typically we look to the causal history of the development of those abilities to see whether or not they are a matter of luck. If we find that the athlete spent no significant time practicing, and that she excelled in every athletic event she dabbled in, we would likely conclude that those abilities are just a matter of luck. But if we find out that she spent hours in the gym practicing and honing her abilities rather than engaging in more immediately enjoyable activities, that would satisfy most of us that her abilities were, at least in part, not a matter of luck.[276] I think the explanation is that putting in the time to practice is, bracketing concerns about determinism, not a matter of luck, but that being born with significant athletic gifts is. Upon finding out that the explanation of the athletic abilities is some non-lucky event, that of deciding to work hard, we treat the presence of athletic abilities as also not a matter of luck.

As a further example, suppose a student were to say that the grade he received on a paper was unfair because both he and a fellow student he knows spent the same amount of time on the paper and exerted the same amount of effort in writing it. When questioned a little further, he reveals that he had worked much less hard earlier in the semester, because of a general laziness or perhaps overconfidence in his academic abilities, than his friend, and so despite the fact that they spent the same amount of time on this paper, one student did worse than another because of how

[276] Only 'in part' because it is not clear from the little story I told that the gym rat was not also blessed with impressive natural athleticism.

much more work he had left himself to do. Finding out that his doing poorly was explained by something that is not a matter of luck, again bracketing issues relating to determinism, is sufficient to show that it is not a matter of luck that he did poorly. To go back to the animal case again, it sounds almost as odd, to me, to say 'It's good luck that dog has four legs' as it is to say 'It is good luck that dog is a mammal.' This is so despite the fact that having four legs is not an essential feature of dogs. Some dogs are born without four legs, and of course some dogs lose a leg or more going through life, while remaining dogs. But when they do have four legs, because the explanation for this will be to cite essential features the dog does have, like the fact that dogs in normal circumstances will have four legs, it is not a matter of luck that the dog has four legs.[277]

I have already presented, in presenting the **Moral Luck Argument**, some reason to think that the two luck principles are false. So I have to say how it is that I think the **Moral Luck Argument** goes wrong. **The Moral Luck Argument** essentially says that starting from an uncontroversial case of luck, a series of small changes that change nothing of moral significance gets us to a case similar in all respects to determinism. Given the nature of the argument there are three ways it could fail. The first is for one or more of the cases mentioned in the argument to be inconceivable or not make sense. The second is for one of the transitions from one case to another to actually mask a significant difference. The third is for the initial case to not actually be an uncontroversial case of luck.

That the **Moral Luck Argument** fails in the third way is implausible. What could be more clearly a matter of luck than someone playing a game like poker against equally skilled players? Certainly people do not deserve praise for winning a poker

[277] Contrast 'having four legs' with 'having three legs'. The first property is explained by an essential features of the dogs, while the second is not. Also, at least to my ears, 'It was back luck for that dog that he was born with three legs' sounds perfectly sensible.

game,[278] and the most natural explanation of why this is true is that victory is just a matter of luck.

I also don't think there is anything incoherent in the cases I have presented. Certainly some of the events described go beyond what we currently know to be physically possible, but there doesn't seem to be any problem with our conceiving of these events taking place. There is a legitimate question as to whether our intuitions about bizarre cases ought to be trusted. Whether this is so depends on why it is that we ought to trust intuitions about normal cases, and that is not an easy thing to explain. Luckily I need not explain it. The only intuitions I appeal to are intuitions about normal cases. I appeal to the intuition that the initial case is one in which victory is a matter of luck, and this is a perfectly normal situation of equally skilled poker players seated around a table. I also appeal to intuitions about what kind of changes can make a difference to how we ought to evaluate whether something is a matter of luck. I claim that change in the stakes ought not make a difference to whether something is a matter of luck, and changes in the stakes of contests is a perfectly normal occurrence. I also claim that changes in one's opponents, as long as equality of skill is maintained should not make a difference to whether something is a matter of luck. What I have not done is pump the reader for intuitions about games of poker played entirely by computers, or games of poker played entirely in one computer. I have not asked the reader what she thinks of these cases, I have told the reader what she ought to think of these cases, given the intuitions I assume we share about the initial case and the significance of the changes between cases. So the metaphysical peculiarity of the last few cases in the **Moral Luck Argument** ought to trouble no one.

[278] Though someone might deserve praise for winning many poker games. But that would just show that the player had higher than average skill at poker, and since equality of skill is being assumed between the players in this case, this issue can be put aside.

Where I think the **Moral Luck Argument** has to go wrong, if it goes wrong, is in one the changes between cases. The first transition is between:

> *Equal Poker:* Jack plays poker for a $1,000 pot. There is nothing odd about the deal, it is a standard poker dealer doing his job as he ought to. The only odd thing about the players is that, while they all know only their cards, they are all equally skilled at calculating the likelihoods of other hands being present in the game given their knowledge of their own hands, at knowing how and when to bluff, etc.. Jack wins $1,000.

And:

> *High Stakes Equal Poker*: In this situation the same factors obtain as in *Equal Poker* except that the pot has changed. Instead of being a game for money, the pot determines which opportunities for careers, social standing, and living environment the children of those at the table will have. Jack wins, earning for his daughter the chance for a great job, high social standing, and great material conditions throughout her life.

Here there is little to worry about. I am tempted to say that *High Stakes Equal Poker* is just as obviously a case of luck as *Equal Poker*.[279]

> The objectionable transition is from *High Stakes Equal Poker* to:
>
> *Absurdly High Stakes Equal Poker*: The situation here is the same as in *High Stakes Equal Poker* except that here every aspect of the lives of the children of the players is determined by who wins, who comes in second, etc. Every hard fact concerning the children of the players is set by who wins the pot. Jack wins, wining for his daughter the same great life, but with every last detail set completely.

Of this transition I said that it represented merely another change in stakes and so should be unobjectionable if the transition from *Equal Poker* to *High Stakes Equal Poker* was. If **Luck as Accidental** is true, however, this transition is objectionable, because for the game to decide every fact about Jack's daughter would be for it to decide what essential properties she has. According to **Luck as Accidental**, for some

[279] I don't think doing so would involve me running afoul of the naturalist's worry about confidence in intuitions about peculiar cases. *High Stakes Equal Poker* could very easily describe something that has happened in the actual world. Imagine some friends playing, for their children, for the final spot at a prestigious educational institution.

property X to hold of some person Y by luck, X must be an accidental property of Y's. So in the transition from *High Stakes Equal Poker* to *Absurdly High Stakes Equal Poker*, some of the things decided by the game stop being matters of luck for Jack's daughter, according to **Luck as Accidental**.

It is still the case that Jack's daughter, lets call her Jane, has all the properties she has because of the game going as it did. It is also still the case that the game going as it did was a matter of luck for both Jack and Jane.[280] The difference between them is it is a matter of luck for Jack that his daughter has the properties she does, but it is not a matter of luck for Jane that she has those properties.[281] This might seem problematic. To make this distinction intuitive, think about a more typical case, this time involving Tom and his daughter Wendy. As before, let us suppose, along with origin essentialists, that it is an essential property of Wendy's that she was conceived from the particular sperm and egg that she was conceived from. Let us also suppose that, given the particular history of that sperm and egg, it follows from this essential property of Wendy's that she was conceived on the day she was conceived. **Luck as Accidental** says that because of this it cannot be a matter of luck for Wendy that she was conceived when she was, and along with **Transfer of Lack of Luck** says that it cannot be a matter of luck for Wendy that she was born around the time she was born. But clearly it is a matter of luck for parents when their children are born. Parents try, sometimes for a week and sometimes for a year, to have children and are thinking perfectly sensibly when they think to themselves that they were lucky to conceive.

4.3.2 Luck in Explanatory Connections?

Transfer of Lack of Luck brings with it both significant indeterminacy and

[280] As I argued before, luck doesn't transfer along lines of causal sufficiency, though I think that lack of luck does.

[281] That one and the same thing could be a matter of luck for one and not the other is not surprising. In football when a player fails to protect the ball and as a result fumbles, it is not a matter of luck for the fumbler. The very same event is a matter of luck for the other team though, as nothing they did led to their good fortune.

significant risk. The indeterminacy comes from the claim that it is sufficient for some action, psychological state or event of X's to not be a matter of luck for X that it be explained by X's essential values. I have not said what is necessary and sufficient for an essential value to explain some action or psychological state or event. I do not intend to give an account of what is necessary and sufficient for explanation. To do would require a work longer than the present one and is not necessary anyway. I am content to rely on the standard kinds of explanations relied on in those fields which seek to explain human action and psychology. Such a reliance is not informative but will do the trick, because any account of explanation I presented which was more or less restrictive than relying on the standard kinds of explanation available in the social sciences would be a controversial account of explanation to start with.

Transfer of Lack of Luck raises another more pressing problem, however. Take some essential value, V, and some non-essential feature of the agent, N. Even if V explains N, given that N is not an essential feature of the agent in question, it is possible for N not to be a feature of that agent at all. So even if V explains N, it was possible for V not to have stood in that relation with N. V could have very well stood in the explanatory relation with some other feature, Q, that is incompatible with N. So the explanatory connection between V and N itself is accidental for the agent. So even if **Luck as Accidental** is true, that the explanatory connection holds could itself be a matter of luck.

There are two questions to consider. One is whether this matters and the other is how, should the compatibilist feel compelled, to argue that it is not a matter of luck that the explanatory connection holds. First to whether it matters. Suppose we are in a position where an agent performed an action, *S*, that is explained by N, and N is in fact explained by V, a value the agent has essentially. Suppose also that we know that V might have explained Q, and that Q could have in turn explained some other action

T. If we accept that the explanatory relations between V and N, and N and *S* are matters of luck, and that V might have explained Q which in turn might have explained *T*, what follows? Well if the argument that Moral Luck is impossible is sound, then the agent cannot deserve credit for performing *S* rather than *T* or for being in psychological state N rather than Q.

If **Transfer of Lack of Luck** is a valid principle, however, the fact that the explanatory connections are themselves a matter of luck does not imply that N or *S* are matters of luck. N is actually explained by V and *S* is actually explained by N, so it would only be the contrastive claims that have to be denied. How serious a problem is this? It does rule out many judgments we are inclined to make. Consider the following example:

> Jack the Pacifist: Jack, a single father of four, is a pacifist. Because of his pacifism he has stopped paying taxes to the United States government so that he does not contribute to war-mongering. The IRS contacts Jack and says that if he does not pay his back taxes that he will be thrown in prison and his children will be split up and go into foster care. Jack faces the prospect of either failing to live up to his duties as a father, or of both failing to live up to his own principles and contributing to the US war machine.

Whether Jack decides to pay his taxes or not he will be guilty of doing something wrong. When people act in such cases we typically temper our judgments of their wrongdoing by making a contrastive claim like 'Yes he abandoned his principles, but he did so rather than keep his family together.' But if determinism is true, then contrastive facts like this are just matters of luck, and so contrastive judgments are inappropriate. Are we stuck blaming Jack for abandoning his principles, without being able to note that he did that rather than abandoning his children? Not in all cases. Suppose that Jack did pay his taxes and that the reason he did was that on this occasion he valued his children more than his cause. Let us also suppose that his

valuing his children either is an essential value of his, or is explained by some other essential value of his. There is a description of his action, that of keeping his family together, under which that action is explained by an essential value of his. So why can't the compatibilist officially eschew all contrastive claims but end up with the ability to praise and blame people in all the cases where contrastive judgments would recommend such praise and blame, simply by shifting the description of the action? If the compatibilist can do this then the loss of contrastive judgments is no serious loss. It requires only a change in the way we speak.

In many cases I think the compatibilist can do this, but not in all. In explaining Jack's action I appealed to the fact that he valued his children more than his cause. This is just another contrastive claim, this time contrasting Jack's commitment to some value rather than contrasting two ways that Jack might have acted. Is this contrastive claim ruled out? It depends. Suppose that Jack's valuing his children is an essential value of his and that his valuing pacifism is not. Then I think the contrastive claim is acceptable. Jack's valuing his children more consists in his valuing his children being an essential value while his valuing pacifism not being one. And it is not an accidental fact that one of those values is essential and the other is not. It is not accidental that Jack is Jack. So contrastive claims about values where one value is essentially held and the other is not seem unobjectionable to me.

A contrastive claim between values that are both essential is problematic. Here we cannot appeal to role these values play in the agent's identity to explain why it is that the agent values one more strongly. If we must appeal to something else, what is it? It is not immediately obvious how to make sense of claims like 'He wants X more than Y'. It does no good to appeal to a pattern of choosing X over Y. That might be good evidence for the claim that 'He wants X more than Y' but it does not tell us what his wanting X more than Y consists in. We might want to talk about the desire for X

having more force than the desire for Y, but absent some account of what such forces are this is just to re-describe the claim, not to explain it. Whatever the answer to this question is, it will simply bring up another question, which is whether the fact or facts which makes it the case that a person valued one thing more than another is just a matter of luck. If it isn't a matter of luck, is it because the agent's valuing X to exactly that extent is an essential feature of her's? Because that is implausible. If it isn't then it seems that we once again lose the contrastive claims.

So if determinism is true then the **Combined View** cannot accommodate contrastive moral judgments in cases where the contrast is between actions or psychological states or events which could both be explained by reference to essential values of the agent. This is not an insignificant failure on the **Combined View's** part. Is there a way to avoid it? I am uncertain, though one way holds some promise. Recall that **Identity Claim** was a substitute for 'taking responsibility' in Fischer and Ravizza's account of moral responsibility. While 'taking responsibility' is insufficient on its own to explain why it is that a mechanism belongs to an agent, there is no reason, once that has been explained by appeal to **Identity Claim** that we cannot make use of the notion of taking responsibility to solve a problem like the one that is before us. Suppose that in Jack's case both values, taking care of his children and pacifism are or are explained by values that Jack has essentially, so that no contrastive claim is appropriate simply on the basis of **Luck as Accidental** and **Transfer of Lack of Luck**. Why can't Jack be responsible for the action because he takes responsibility for the action?

If his taking responsibility for this action were itself an action he was responsible on the **Combined View** then it seems as though there is no obstacle. His deserving to be praised or blamed for the contrastive fact would be something like a case of derived desert. In taking responsibility Jack performed an action something

like a promise. In this case he opened himself up to others to praise or blame for his action, thereby making it fair for others to do so. Why couldn't such an account save contrastive judgments from the threat of determinism? To defend such a claim fully would require taking a much closer look at the ways in which derived desert can arise. This would involve giving an account of promising, among other things, and the giving of such an account would again involve writing a longer essay than the present one. I offer this suggestion as a way compatibilists might hope to save contrastive judgments, though for my part I am comfortable letting them go for the sake of the **Combined View**.

4.4 The Combined View and Manipulation Arguments

Recently the most popular kind of argument for incompatibilism has not been a transfer argument, and has not relied on **PAP**. The most popular kind of argument for incompatibilism is the **Manipulation Argument**.[282] Essentially all **Manipulation Arguments** are committed to two claims. The first is that if some person X, performed some action Y because they were manipulated into doing so, then X is not responsible for doing Y. The second is that if determinism obtains then X's relation to Y is morally indistinguishable from what it would be if X were manipulated. Together these claims imply that if determinism obtains then X is not responsible for doing Y.

To deny the first claim, that manipulation rules out moral responsibility, is to pursue a hard-line defense of compatibilism, while to deny the second, that manipulation is morally indistinguishable from determinism, is to pursue a soft-line response. The **Combined View** provides some reason to pursue both defenses. Which defense it recommends is going to depend on the kind of manipulation in question.

[282] For some central discussions of the **Manipulation Argument** see Fischer(2004), Kane (1996), Mele (1995, 2005, 2006), McKenna (2004, 2008) and Pereboom (2001, 2008)

Consider a relatively straightforward story about manipulation:

> *Mad Confectioner:* William is a confectioner who was very worried about the popularity of low-carb, low-sugar diets and what these new healthy habits would do to his business. You see William used to be a unscrupulous neurobiologist and invested all the ill-gotten gains of his profession into a confectionery shop, so he stand to lose everything. So William concocted and released into the water a chemical compound which produces in people an overpowering and almost insatiable desire to consume sweets. Everyone in the area then broke their diets and started frequenting his shops. The power of the implanted desire is so strong however that the locals spent all their money on sweets and, once the money ran out they began to steal from William, who quickly went out of business.

In this case William manipulated his neighbors by implanting a desire in them. And it seems clear that William's neighbors are not responsible for their thieving, or for putting William out of business by thieving. The **Combined View** agrees with the intuitive result. The explanation for why William's neighbors stole from him does not involve their values. In this story a very strong desire was simply added to their existing set of psychological states and dispositions. The desire was strong enough that none of the pre-existing values held by William's neighbors could have led them to do anything other than buy or steal sweets. Given the insensitivity of their thieving to their values, those values do not explain their behavior in any significant way.

So for any **Manipulation Argument** that starts with examples of manipulation like *Mad Confectioner* a defender of the **Combined View** will have to take a soft-line approach.. To see how a soft-line response works we need an example of a **Manipulation Argument**. Just as with the **Moral Luck Argument** we start with an uncontroversial case, like *Mad Confectioner*, and move through a series of cases which are supposed to differ only in morally irrelevant ways. The next case is:

> *Grand Mad Confectioner*: Just as before William is a confectioner with a past in neurobiology. In this case, however, he has far greater resources at his disposal and instead of setting up shop in a town and

> then drugging its inhabitants, he decides to create his own customers. So William uses his expertise in biology and neurobiology to create human beings in such a way that he can guarantee that they will have certain beliefs, desires and dispositions by the time they are old enough to make their own decisions about how to spend their money. They are created so as to have stable, reliable and extreme desires to consume sweets. These desires cohere perfectly with all their other mental states and dispositions, or at least as well as a normal person's ends cohere with those states and dispositions. As before these desires lead to bankruptcy of William's creations and then to their stealing from him.

Here the transition just involves the manipulation becoming more distant. Instead of taking already formed human beings and implanting a single desire, William creates the human beings whose behavior he wishes to guarantee much earlier. But they are still guaranteed to end up stealing from William. The next transition is to:

> *Capitalist Confectioner:* William is again a maker of sweets, but this time what makes his neighbors have their voracious appetites for sweets is consumer culture more generally. They have the desires they have, the character they have, and the beliefs they have (all of which fit together nicely) because of the effects of corporate advertising, an education system hostage to corporate interests, and low wage/long hour jobs that prevent their parents from correcting for these influences. These desires are just as strong as in the earlier examples, and they also lead to bankruptcy and theft.

Here no one in particular is responsible for guaranteeing the behavior of William's neighbors, but their behavior is nonetheless guaranteed, and just as completely as in the *Grand Mad Confectioner* case. The final transition is from *Capitalist Confectioner* to what would be the case if determinism obtained. If general social causes present a problem for moral responsibility it is hard to see why general physical causes could fail to.

With this sequence of *Confectioner* cases you could pursue a soft-line response in objecting to the transition from *Mad Confectioner* to *Grand Mad Confectioner*, *Grand Mad Confectioner* to *Capitalist Confectioner* or from *Capitalist Confectioner* to the ordinary case of determinism. The best place for the defender of the **Combined**

View to object is in the first transition. What made William's neighbors not responsible for their thieving in *Mad Confectioner* was the fact that their actions did not express their true values. But in *Grand Mad Confectioner* their actions do express their values. In that case William gave his neighbors values that cohered perfectly with the overwhelming desire for sweets. What is going on in *Grand Mad Confectioner* is that William lives with a bunch of people who have made the pursuit of sweets their dominant practical end. Their actions express the value they place on eating sweets, and without any more information about their psychological state, there is no reason to think either that they fail to properly appreciate the relevant non-normative facts or that they fail to appreciate the reasons which supervene on those non-normative facts. If it were the case that they were normatively incapacitated in this way, then the **Combined View** would recommend resisting the next transition, because corporate advertising does not render us incapable of appreciating reasons or acting on them, though it might make it less likely that we will.

Things would be different if the initial case of manipulation were, like *Grand Mad Confectioner*, a case of global manipulation of the agents involved. If that were the case then the **Combined View** would recommend the hard-line approach, and say that in that case manipulation was not incompatible with moral responsibility. In fact, on one way of looking at it, the **Combined View** is already committed to a hard-line response, in that it implies that manipulated agents can be responsible for what they do.[283] Simply asserting that manipulation is not necessarily incompatible with moral responsibility is easy enough, but what a view like the **Combined View** does is give an explanation of why that is the case.

[283] The possibility here is broad metaphysical possibility, as it is not clear that any of the cases of global manipulation in the literature are nomically possible.

Bibliography

Alston, William (1988) "The Deontological Conception of Epistemic Justification," *Philosophical Perspectives* v. 2 pgs.115-52

Arpaly, Nomy (2006) *Merit, Meaning, and Human Bondage: An Essay on Free Will* Princeton: Princeton University Press
_____, (2003) *Unprincipled Virtue*: *An Inquiry into Moral Agency* OUP:Oxford

Ayer, Alfred Jules (1954) "Freedom and Necessity" *Philosophical Essays* ed. Steven Cahn London: MacMillan

Bennett, Jonathan (1980) "Accountability" in *Philosophical Subjects*, Zak Van Straaten, ed. Oxford: Clarendon Press

Bratman, Michael (2007) *Structures of Agency* OUP:Oxford

Buss, Sarah and Overton, Lee (2002) *Contours of Agency: Essays on Themes from Harry Frankfurt* MIT Press: Cambridge, MA

Chisholm, Roderick (1982). "Human Freedom and the Self," in Watson (2003)

Cohen, G.A. (1989) "The Currency of Egalitarian Justice" *Ethics* vol. 99 n. 4

Curley, E.M. (1975) "Descartes, Spinoza, and the Ethics of Belief." In *Spinoza: Essays in Interpretation*, ed. Maurice Mandelbaum and Eugene Freeman, LaSalle, Ill.: Open Court Publishing

Damasio, Antonio (1999) *The Feeling of what Happens: Body and Emotion in the Making of Consciousness*, New York: Harcourt Brace and Co.

De Sousa, Ronald (1987) *The Rationality of Emotion*, Cambridge, MA: MIT Press

Della Rocca, Michael (1998) "Frankfurt, Fischer and Flickers" in *Nous* v. 32 n.1 pgs. 99-105

Dewey, John (1891) *Outlines of a Critical Theory of Ethics* (New York: Hillary House, 1957),

Dretske, Fred (2000) "Entitlement: Epistemic Rights without Epistemic Duties?" *Philosophy and Phenomenological Research* v. LX n. 3 pgs. 591-606

Dworkin, Ronald (2000) *Sovereign Virtue: The Theory and Practice of Equality* Cambridge, MA: Harvard University Press

Feldman, Fred (2004) *Pleasure and the Good Life: On the Nature, Varieties, and Plausibility of Hedonism* Oxford University Press: Oxford

Feldman, Richard (2000) "The Ethics of Belief" *Philosophy and Phenomenological Research* v. LX n. 3 pgs.667-96

Fischer, John Martin (1995) *The Metaphysics of Free Will: An Essay on Control* Blackwell
_____, (2004) "Responsibility and Manipulation" *Journal of Ethics* 8: pgs. 145-177

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

Frankfurt, Harry (1971) "Freedom of the Will and the Concept of the Person" in *Journal of Philosophy* v. 68 n. 1
_____, (1969) "Alternative Possibilities and Moral Responsibility" in *Journal of Philosophy* v. 66 n. 3
_____, (2002) "Reply to J. David Velleman" in Buss and Overton(2002)

Ginet, Carl (1966) "Might We Have No Choice" in *Freedom and Determinism* ed. Keith Lehrer. Random House
_____, (2006) "Working with Fischer and Ravizza's Account of Moral Responsibility" *Journal of Ethics* v. 10 n. 3 pgs. 229-253

Glover, Jonathan (1970) *Responsibility* (New York: Humanities Press).

Greco, J. (1995) "A Second Paradox Concerning Responsibility and Luck", *Metaphilosophy,* 26 pgs. 81-96.

Haji, Ishtiyaque (1998) *Moral Appraisability* OUP: Oxford
_____, (2002) *Deontic Morality and Control* Cambridge University Press

Hobart (1934) "Freedom of the Will as Requiring Determinism and Being Inconceivable Without It" in *Mind* v. 143 n. 169

Hobbes, Thomas (1994) *Leviathan* ed. Edwin Curley Hackett Publishing

Hurley, Susan (1999) "Responsibility, Reason and Irrelevant Alternatives" *Philosophy and Public Affairs* v. 28 n. 3 pgs. 205-241

James, William (1884). "What is an Emotion?" *Mind* v. 9 pgs. 188–205.

Kane, Robert (1996) *The Significance of Free Will* OUP: Oxford

Lackey, Jennifer (2008) "What Luck is Not" *Australasian Journal of Philosophy* v. 86 n. 2

Latus, A. (2000) "Moral and Epistemic Luck", *Journal of Philosophical Research,* 25 pgs. 149-172

Lehrer, Keith, (1968) “Cans without Ifs,” *Analysis*, 29 pgs. 29–32.

Lewis, David (1981) “Are we Free to Break the Laws” Theoria, v. 47 n.3 pgs.113–121

Levy, Neil (2008) “Bad Luck Once Again” in *Philosophical and Phenomenological Research* v. 77 n. 3
_____, (2009a) “Luck and History Sensitive Compatibilism” in *The Philosophical Quarterly* v. 59 n. 235
_____, (2009b) “What and Where the Luck Is: A Reply to Jennifer Lackey” in *Australasian Journal of Philosophy* v. 87 n. 3

MacGill, Kevin (1997) *Freedom and Experience: Self-Determination without Illusions* New York: St. Martins Press

Moore, G.E. (1993) “Free Will” in *Principia Ethica* (Cambridge University Press: Cambridge)

McKenna, Michael (1998) “The Limits of Evil and the Role of Moral Address: A Defense of Strawsonian Compatibilism.” *Journal of Ethic* v..2 pgs. 123–142
_____, (2005) “Where Frankfurt and Strawson Meet.” *Midwest Studies in Philosophy* 29
_____, (2001) “John Martin Fischer and Mark Ravizza's *Responsibility & Control*,” Review. *Journal of Philosophy*, XCVIII (2): 93–100
_____, (2009) “Compatibilism” in *Stanford Encyclopedia of Philosophy*
_____, (2004) “Responsibility and Globally Manipulated Agents” *Philosophical Topics* 32: 169-192
_____, (2008) “A Hard-Line Reply to Pereboom’s four case Manipulation Argument” *Philosophy and Phenomenological Research* v. 71 n. 1

McKenna, Michael and Widerker, David (2003) *Moral Responsibility and Alternative Possibilities* Ashgate Publishing Company: Burlington VT

Mele, Alfred (2006) *Free Will and Luck*, Oxford University Press: Oxford
_____, (2000) “Reactive Attitudes, Reactivity and Omissions” *Philosophy and Phenomenological Research* v.61 pgs. 447-452
_____, (1995) *Autonomous Agents* Oxford: Oxford University Press
_____, (2005) “A Critique of Pereboom’s Four Case Argument for Incompatibilism” *Analysis* 65:75-80

Nagel (1993) “Moral Luck” in Statman (1993)

Nelkin, Dana (2009) “Responsibility and Rational Abilities: Defending an Asymmetrical View” *Philosophical Explorations* v. 12 n. 2 pgs. 151-165

Neu, Jerome (2000) *A Tear is an Intellectual Thing: the Meaning of Emotions*,

Oxford, New York: Oxford University Press

Nowell-Smith (1948) Peter "Freewill and Moral Responsibility" *Mind* v. 57 JA pgs. 45-61

Nozick, R. (1974) *Anarchy, State, and Utopia*, New York: Basic Books

Nussbaum, Martha (2001) *Upheavals of Thought: The Intelligence of Emotions*, Cambridge: Cambridge University Press

Pereboom, Derk (2001) *Living Without Free Will*, Cambridge University Press
_____, (2003) "Source Incompatibilism and Alternative Possibilities" in McKenna and Widerker
_____, (2008) "A Hard-Line Reply to the Multiple Case Manipulation Argument" *Philosophical and Phenomenological Research* v. 77 n.1

Plantinga, Alvin (1993) *Warrant: The Current Debate* (New York: OUP)

Pojman, Louis (1993) "Believing, Willing and the Ethics of Belief" *The Theory of Knowledge* ed. L Pojman (Belmont CA: Wadsworth) 525-43

Prinz, Jesse (2004) *Gut Reactions: a Perceptual Theory of Emotion*, Oxford: Oxford University Press

Rawls, John (1971) *A Theory of Justice* Cambridge, MA: Harvard University Press

Rhees, Martin (2000) *Just Six Numbers* Basic Books: New York

Rorty, Amélie (ed.) (1980) *Explaining Emotions* 103–126. Los Angeles: University of California Press.

Russell, Paul. (1992) "Strawson's Way of Naturalizing Responsibility." *Ethics* 102: 287–302
_____, (2000) "Review *Responsibility and Control*" *Canadian Journal of Philosophy* v.32 n.4 pgs. 587-606

Scanlon, T. M. (1998) *What We Owe To Each Other* Cambridge, MA: Harvard University Press
_____, (2008) *Moral Dimensions: Permissibility, Meaning, Blame*, Harvard University Press
_____, (1988) "The Significance of Choice." In *The Tanner Lectures on Human Values*, Vol. 8 Salt Lake City: University of Utah Press

Schlick, Moritz (1939) *Problems of Ethics* Prentice Hall: New York

Scott-Kakures, Dion (1994) "On Belief and Captivity of the Will." *Philosophy and Phenomenological Research* 54: 77-103.

Shabo, Seth (2005) “Fischer and Ravizza on History and Ownership” *Philosophical Explorations* v. 8 n. 2, 108-114

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

Solomon, Robert (2004) “Emotions, Thoughts and Feelings: Emotions as Engagements with the World” in *Thinking About Feeling: Contemporary Philosophers on Emotions* ed. Solomon, Robert OUP: New York

Statman, Daniel (1993) *Moral Luck* SUNY: Albany

Stern, Lawrence, (1974) “Freedom, Blame, and the Moral Community.” *The Journal of Philosophy* 71: 72–84

Stocker, Michael (1979) “Desiring the Bad: An Essay in Moral Psychology” in *Journal of Philosophy* v. 76 n.12

Strawson, Galen (1994) “The Impossibility of Moral Responsibility” *Philosophical Studies* v.75 n.1-2 (1994)

Strawson, Peter (1974) “Freedom and Resentment.” in *Freedom and Resentment and other Essays*. London: Methuen
_____, (1985) *Skepticism and Naturalism: Some Varieties*. New York: Columbia University Press

Van Inwagen, Peter (1980) "The Incompatibility of Responsibility and Determinism," in M. Bradie and M. Brand, eds., *Bowling Green Studies in Applied Philosophy*, Vol. 2 Bowling Green, Oh.: Bowling Green State University Press pp. 30-7
_____, (1983) *An Essay on Free Will* OUP: Oxford

Velleman, J David (1999) “Love as Moral Emotion” in *Ethics* 109: 338-374
____, (1989) *Practical Reflection*, Princeton: Princeton University Press
____, (2002) “Identification and Identity” in Buss and Overton (2002)

Wallace, R. J. (1994)*Responsibility and the Moral Sentiments* Cambridge, MA: Harvard University Press

Watson Gary, (1987) “Responsibility and the Limits of Evil” in *Responsibility, Character, and the Emotions*. Ed. Ferdinand Schoeman New York: Cambridge University Press
_____, (1975) “Free Agency” in *Journal of Philosophy* 72: April 202-220
_____, (1977) “Skepticism about Weakness of the Will” *Philosophical Review* 86:3
_____, (1987) “Free Will and Free Action” *Mind* 96 145-172
_____, (1996) “Two Faces of Responsibility” *Philosophical Topics* 24/2 227-248
_____, (2003) *Free Will 2nd Edition* OUP: Oxford

Williams, Bernard (1973) “Deciding to Believe” in *Problems of the Self* Cambridge: Cambridge University Press

Wolf, Susan (1990) *Freedom Within Reason*, OUP:Oxford

Zimmerman, David (2001) “Thinking with Your Hypothalmus: Reflections on a Cognitive Role for the Reactive Emotions” in *Philosophy and Phenomenological Research* v. 63 n. 3 pg. 521-541

Zimmerman, Michael (1988) *An Essay on Moral Responsibility*. Totowa, NJ: Roman and Littlefield
_____, (1987) Luck and Moral Responsibility”, *Ethics*, 97: 374-386