PERVASIVENESS OF SELF: A CRITIQUE OF P.F. STRAWSON'S REACTIVE THEORY OF RESPONSIBILITY

by

ANDREA JACQUELINE SCOTLAND

B.A. (Hon.), Simon Fraser University, 1999

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

Department of Philosophy

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 2001

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Philosophy

The University of British Columbia
Vancouver, Canada

Date Oct. 12, 2001

ABSTRACT

In this thesis, I argue that P.F. Strawson seriously *underestimates the controversial nature* of the beliefs and attitudes of which the retributive reactive attitudes (RAs) often involve. Although he acknowledges that the RAs involve a "seamy side," he fails to admit they *frequently* commit the reactive person to psychological, if not metaphysical, beliefs that *violate principles of impartiality* and *rationality*. As we shall, this *is* important in the *dialectical context* of "Freedom and Resentment", because Strawson's goal of *reconciling* the compatibilists and incompatibilists about moral responsibility requires the RAs to be free of such controversial presuppositions. I argue that because more modest versions of "seamy" retributive RAs *are* grounded in false and egoistic beliefs, the incompatibilist will remain skeptical that the gap in consequentialist compatibilism can be filled by the *non*-metaphysical fact of our "*natural proneness*" to take up the reactive stance.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

> "It has been said that the highest praise of God consists in the denial of Him by the atheist, who finds creation so perfect that he can dispense with a creator."
>
> — Marcel Proust

The love and assistance of the following people have given me the ability to write this thesis, and, more importantly, to find "creation so perfect": John Christianson, Kari Coleman, the Dick family (especially my stepmother Marlies), Professor Jim Dybikowski, Mary Graydon, Professor Phil Hanson, Bill Hay, Professor Ray Jennings, Joanne Miller, Professor Paul Russell, Jeff and Deanna Scotland, Miller Scotland-Sobolewski, Joe Sheridan, Sondra Sterling, and Tanya Wulff. I am particularly indebted to Mahri Mackenzie, Earl Marsh, Val and Gordon Scotland, and Jacqueline and Ed Unger for their unbounded parental love, and to Debbie B., Dawn D., Wendy L., Lisa P. and other girlfriends of Bill for (my Higher Power) *their* inspiration. Finally, the "highest praise" to the blissful reprieve of Wendy Miko and Judith Stapleton, the technical genius of Nicole Friedrich and Chuck Robinson, the untiring devotion of Saint Jimbo (Jim Cullina), and, especially, the love (and enduring patience) of Steve Sobolewski.

DEDICATION

Mon raison d'etre Nathanael and Donovan Dick.

INTRODUCTION

> What sort of 'true' moral responsibility is being said to be both impossible and widely believed in? An old story is helpful in clarifying this question. This is the story of heaven and hell. As I understand it, true moral responsibility is responsibility of such a kind that, if we have it, then it *makes sense*, at least, to suppose that it could be just to punish some of us with (eternal) torment in hell and reward others with (eternal) bliss in heaven (Galen Strawson, "The Impossibility of Moral Responsibility" 9).

One does not have to believe in heaven and hell to be among those for whom the existence of moral responsibility "makes sense." In fact, we may ordinarily *assume* holding persons morally responsible, for at least some of their actions, at least some of the time, is appropriate. While most people *do not* hold roses morally responsible for pricks of thorns nor snakes for venomous bites, many *do* hold persons morally responsible, at least sometimes, for things like murder or neglecting a child. Although we acknowledge persons commit wrongs for which they are *not* responsible, such as those committed accidentally, many suppose there *are* others, whether Nazi war crimes, corporate tax evasion or sexual harassment, for which persons *should* be held morally responsible.

But, although many in society *assume* we ought to hold persons morally responsible at least sometimes, under what conditions (if any) *does it* "make sense" to do so? If the terrain were not hazardous enough, the question of human freedom and moral responsibility becomes particularly daunting when we ponder the truth of metaphysical determinism: *Is* moral responsibility *compatible* with a causally determined world? *If* all our choices, decisions and actions are necessitated by preceding causes, how *can* it be right to blame or

punish persons for the wrongs they commit? Is it appropriate for wrongdoers, and those holding them to account, to experience negative moral emotions such as guilt, or resentment — if everything (including human action) *is* determined by causal antecedents?[1] What if many, if not all, ordinary assumptions about moral responsibility are mistaken? We may be right, children, the mentally ill, the compelled and coerced — to say nothing of rose bushes and slithering snakes — should *not* be held morally responsible. But, contrary to what many in society think, the truth of metaphysical determinism may mean roses pricking fingers and snakes that bite, may share important similarities with parents abusing children and mercenaries that fight. A causally determined world could mean everything, from the stings of nettles to human slavery, *is* caused by prior events over which persons possess little to no control. Consequently, many, perhaps all, beliefs presupposed by the practice of attributing moral responsibility may turn out to be false.

Moreover, over and beyond the problems involved with the compatibility of moral responsibility and determinism, providing a satisfactory account that coheres with our sense of justice also proves elusive. Historically, full personhood or moral responsibility for actions was thought to require such odd prerequisites as being male, the ownership of private property, the possession of a human soul, free will, even Christian baptism (Lloyd). Today, in spite of metaphysical and theological controversies such as grace and predestination, many continue to support the idea that moral responsibility requires free will (if not private property, maleness, or Christian baptism). Attempts are made to strip free will of it's obscure metaphysical and theological pretensions as being morally responsible is said to require 'free choice,' 'free action,' 'alternative possibilities' or the 'ability to act otherwise' (Kane, Van Inwagen). Although the association of moral responsibility with free will *continues* to play a

strong roll in the Western psyche, many reject the notion as superstitious and unsubstantiated by the human and social sciences. Rather than grounding moral responsibility in free will and a conception of persons as the "ultimate, buck-stopping originator(s)" of their actions (Galen Strawson, *Freedom and Belief* 26), there are others who argue moral responsibility requires certain sorts of "control" over actions, and the completion of a series of antecedent deliberative processes (Fisher and Ravizza).

Here, however, we focus on yet another alternative to the free will legacy. According to this perspective, rather than deliberation and control, appropriate attributions of praise and blame, punishment and reward presuppose little more than beliefs about another's "*intentions*" (P.F. Strawson, "Freedom and Resentment"). Rather than such dubious criteria as owning private property, or belonging to the right race or gender, the nature of a person's intentions *may* (somehow) represent something *morally meaningful* about them, such as the expression of their higher-order commitments (Frankfurt, "Freedom of the Will"). For some of us this may sound right. We may, in fact, blame or hold perpetrators to account upon discovering the presence of intentions to harm, or the lack of good intentions or care. But, what *makes* a person's intentions morally significant? What if the *source* of a person's intentions raises doubts about their importance to moral responsibility?

Versions of the following problem pose a perennial haunt for theorists of all stripes in the moral responsibility debate. Suppose we find at least some of the proceeding makes sense. We may agree, for instance, people *do* ordinarily attribute moral responsibility or experience resentment towards those who harm us, upon coming to believe the perpetrator freely chose or intends to harm us. But, we may also agree part of what it means to act freely or intend to inflict harm *is wanting* or *desiring* that one's actions inflict harm. After all, it is

hard to imagine someone who freely chooses or intends to inflict harm, who *is* morally responsible, who does *not want* to inflict the harm they are morally responsible for inflicting. It just doesn't seem right to hold someone morally responsible for harming us if we discover the perpetrator did *not* want to harm us. It seems whether the solution to the moral responsibility debate involves free will, control, *or* intentional action, all seem to involve the nature of human wants or preferences.

But, if the nature of wants or desires *is* central to any adequate account of moral responsibility, how do we know someone does, in fact, wants to act the way they do? Some theorists argue persons *do* want to act the way they do, unless they are *physically* forced to do so by some *external* force (Hobbes). In other words, according to this minimalist position, persons *want* to do pretty much *all* they *actually do*, (including inflict harms). Unless pushed, shoved or are otherwise forced by some external physical source, we ought to hold all wrongdoers morally responsible, blame, even punish, them. But, clearly, this won't do. There are all sorts of actions committed (or omitted) without the involvement of a push or a shove (or any other external physical force or impediment) for which persons do *not* seem morally responsible. For instance, the mentally ill and children act without being physically forced. It does *not*, however, seem right to hold them morally responsible.

If "wanting" or "desiring" to act is defined minimally as acting in the absence of physical coercion or impediment, then, sufferers of childhood abuse, addiction, brainwashing, psychological coercion, desperation, hypnotism and so on, *would* be held morally responsible, blamed, even punished for most (if not all) their wrongs. If moral responsibility involves such things as free will, control, or intentional action, but these also involve personal preferences or desires, then any adequate account of moral responsibility

may face the problem of not only identifying, but also evaluating the source of preferences and desires. After all, acquiring preferences or desires may involve such morally dubious sources as propaganda, hypnotism, addiction or fear. Praising someone for donating blood seems odd — if the *desire* to donate is, in some sense, *implanted* by something or someone else, e.g., a benevolent hypnotist or brainwashing (Frankfurt, "Freedom of the Will"). Similarly, if the *source* of a person's desire or intentions to commit such things as violence against women, is also, in some meaningful sense, "implanted," then perhaps *these* perpetrators should *not* be held morally responsible. A shocking conclusion indeed. Perhaps parental cruelty, life in the ghetto, or even the media, can be said to "implant" desires or cultivate intentions. If it is *not* appropriate to blame persons for wrongs they commit as the result of "implanted" desires, and, such forces as racial and gender stereotypes, child abuse, addiction, learning disabilities, neural-chemical imbalances, inadequate nutrition, low self esteem, economic inequalities, and so on, *do*, in some meaningful sense, "implant" desires, perhaps most persons (if not all) should *not* be held morally responsible or blamed for their wrongs. An even more shocking conclusion. When *are* we morally responsible for the actions resulting from our desires? What kind of moral education, family, role models, nutrition, esteem *is* necessary for the formation of morally relevant human desires? Are certain political rights and economic advantages required? Unless we develop historical criteria by which the *source* of a person's preferences and desires *can* be identified and evaluated, concerns remain about the prospects of a satisfactory account of moral responsibility. Holding persons morally responsible for actions they "freely choose" or "intend" to commit — particularly when blame, the negative moral attitudes and punishment are involved — may be inappropriate (if not unjust). Whether a satisfactory account of moral

responsibility involves either the notion of free will *or* intentional action, the problem of the *source* of another's desires looms large.[2]

But what about the desire *to hold others morally responsible itself*? If the source of desires to act *is* important to appropriate attributions of moral responsibility, perhaps the source of desires to attribute moral responsibility is *also* important. For instance, we may want to blame bad drivers for cutting us off in rush hour traffic. But, we may tend to blame them when *we* are late or worried about arriving on time. We may seldom wag our fingers or utter a "tsk tsk" in response to rude drivers when *we* are on time and relaxed. Similarly, we may find ourselves wanting accountability from those who harm children. But, our feelings of moral indignation may markedly increase if the children harmed are *our own*. And, perhaps the fervent *desire* to hold persons like Bill Clinton morally responsible for *his* sexual crimes stem, at least in part, from the guilt or feelings of inadequacy of those who most zealously sought accountability. If the source of desires to act *is* important to appropriate attributions of moral responsibility, perhaps the source of desires to attribute moral responsibility is *also* important. Our discussions here may lead to the conclusion that, accounting for the source of *our desires to hold another* morally responsible, may prove as important to moral responsibility as accounting for the source of *another's desires to act.* Developing an account of moral responsibility that *is* compatible with the truth of metaphysical determinism *and* principles of justice is, then, profoundly important and highly contentious — if not out right baffling.

Philosophers have traditionally distinguished themselves in response to the former, namely, difficulties raised by the free will determinism debate in one of two ways: Compatibilists believe the practice of moral responsibility is, or can be "compatible" or

appropriate in a world governed by deterministic laws. Incompatabilists believe the practice is not.

Compatibilists such as P.F. Strawson[3] seek to vindicate the practice of moral responsibility in part by arguing the conditions constraining appropriate *retributive reactive attitudes* (hereafter, retributive RAs)[4] do *not* involve false or unjustified beliefs about the world. In "Freedom and Resentment" and elsewhere Strawson argues, more specifically, that adopting retributive RAs like resentment and indignation does *not*, at least ordinarily, involve believing anything controversial about the nature of human action more generally. For instance, many experience feelings of moral indignation towards highjackers for collapsing the World Trade Center and civilian death. Others experience similar feelings towards American politicians for foreign policy and civilian death. Those adopting the indignant stance may believe *many* things about the world and the people in it — *including* theoretical beliefs about human nature, the undetermined nature of human action, free will, and so on. Strawson's point, however, is that, ordinarily the person's experience of moral indignation is *independent* of (any) controversial theoretical beliefs involving the falsity of determinism. Rather than metaphysical beliefs about free will, the undetermined nature of human action, immortal souls, and so on, the retributive attitudes involve nothing more controversial than our "sense of sympathy and common humanity" (P.F. Strawson, *Strawson Replies* 266), and beliefs about the quality of a perpetrator's intentions. Grounded in deep *care* and *concern* for *humanity*, Strawson suggests indignation towards highjackers (or American politicians) are about the *wrongdoers' will* or *intentions* towards their victims. Do hijackers (or American politicians) display adequate concern for the interests of *all* humanity? If not, are they burdened by standard excusing and exempting conditions such as insanity or coercion?

Strawson's strategy involves persuading us that the retributive RAs typically involve answers to these philosophically lean questions — and *not* presuppositions about determinism or the springs of human action more generally.[5] If he is right, the practice of moral responsibility may be compatible with a causally determined world.

In this thesis, I argue that this reconciliationist strategy is likely to fail — even if the incompatibilist can be convinced that the reactive stance does *not* presuppose robust metaphysical beliefs. In this thesis, I argue that Strawson seriously *underestimates the controversial nature* of the beliefs and attitudes, which the retributive RAs often involve.[6] I argue that *frequently* they commit the reactive person to psychological, if not metaphysical, beliefs that *violate principles of impartiality*, if not *epistemic rationality*. As we shall see, this *is* important in the *dialectical context* of "Freedom and Resentment," because Strawson wishes to *reconcile* the compatibilists ("optimists") and incompatibilists ("pessimists"). His strategy is to persuade the pessimists to give up their "panicky metaphysics" (25) of free will, agent-causes, noumenal selves and the like, in return for a major concession from traditional optimists of the sort that place exclusive emphasis on the *consequentialist* aspects of moral responsibility. While Strawson agrees with incompatibilists that this emphasis on expediency drains moral practice of something crucial, he argues against them that the lacuna cannot be filled at a lower philosophical price. In short, Strawson hopes to reconcile the disputants within the moral responsibility debate by persuading incompatibilists that *the* "vital thing" (23) missing from the optimist account *can* be restored without controversial metaphysical beliefs.

It seems to me, however, Strawson's reconciliationist strategy of détente is likely to succeed only if *the feature* he reintroduces into the responsibility arena *is* truly *attractive* to

the incompatibilist — otherwise the incompatibilist will *fail* to be convinced that the gap in consequentialist compatibilism can be filled by *non*-metaphysical facts like the "natural proneness" to take up the reactive stance. Strawson suggests incompatibilists suffer "*emotional shock*" (21, italics mine) towards consequentialist compatibilism for neglecting that which makes us recognizably human (23), the *care and concern* for the goodwill of others towards us and *all* humanity (15). Strawson must bait his hook, then, with something that must be understood *as* "vital" to the practice of moral responsibility, that salient phenomenology which itself expresses the care and concern for our common humanity. If, on the other hand, Strawson's bait seems seriously *tainted* in some obvious fashion, incompatibilists will lose their appetite and become skeptical about the prospects of reconciliation with the optimists, based on natural facts.

Strawson's idea is to ease the incompatibilists "emotional shock" and ground the vitality of the practice of responsibility in the *emotional richness* of the RAs, instead of controversial metaphysics. In this thesis I argue that this reconciliationist strategy is likely to fail — even if the incompatibilist can be convinced that the reactive stance does *not* presuppose robust metaphysical beliefs. I wish to show that there *are* important features of the retributive stance which are likely to encourage incompatibilists to regard them as "tainted goods" or excessively "seamy" — but not *the* "vital thing" missing from the optimist account.

What is the "seamy side" of the retributive RAs? Strawson admits that the psychological sciences are, of late, becoming increasingly aware that the retributive RAs sometimes involve the "self-deception of the ambiguous and the shady" such as "unconscious sadism and the rest." He argues that although the *complete* elimination of the

network of RAs from our lives *is* an empirical impossibility, we may, nevertheless, want to redirect and modify our retributive RAs in light of these studies (25). But, according to Strawson, even if we *could* eliminate the RAs from our lives, these studies should *not* "increase the difficulty" of acknowledging that the "humanity" of the RAs themselves are the "only possibility of reconciling" the pessimists and the optimists (24). According to Strawson then, acknowledging the "seamy side" of the RAs should *not deter reconciliation*, for it is "an exaggerated horror, itself suspect, which would make us unable to acknowledge the facts because of the seamy side of the facts" (25).

But, contrary to what Strawson might think, I argue he seriously underestimates the *dialectical import* of the "seamy side" of the retributive attitudes *for* he neglects the seamy side's *more modest versions*. When incompatibilists consider *both* the "shady" *and* more modest versions of the "seamy side" of the retributive RAs, they will lose their appetite and fail to trade in their metaphysics as Strawson's reconciliationist strategy requires.

First, I argue that Strawsonian minimalism, that is, the thesis that *uncontroversial psychological beliefs and attitudes* about the intentions and histories of others, *often* fails to explain certain of the temporal and qualitative features — sometimes even the very existence — of the retributive RAs. I argue further that, more often than Strawson thinks, features of the retributive RAs of resentment and indignation *can* be explained only in terms of certain *reflexive psychological beliefs and attitudes* that the reactor is *unaware of*. We will see, that in an alarmingly wide range of cases, positing these unacknowledged self-directed beliefs and attitudes fills an *explanatory lacuna* in *Strawson's account*. What makes us right to suspect the "seamy" side of the RAs is more problematic than Strawson supposes? Positing such dubious reflexive beliefs provides the *best explanation* for the Duration and Intensity of

many retributive RAs, for how and whether they Respond to historical pleas, and for whether they come to Exist in the first place. I call this the *D.I.R.E.* dimension of RA formation.

I argue in some detail that accounting for the D.I.R.E. nature of a person's other-directed retributive RAs of resentment and indignation often requires that we posit, in the reactor herself, the existence of unacknowledged *reflexive* beliefs and attitudes about a *plurality* of possible factors, such as the person's own *moral guilt*, her *ethnic* and *familial* attachments, or her own *role as a victim* in similar situations in the past. If I am right, and reflexive beliefs about the *reactor's own* history and intentions play a critical role in explaining his or her retributive RAs towards *others*, then, contrary to Strawson, deep doubts *do* arise about the Strawsonian reconciliationist project; if the so called "moral" attitudes we adopt *towards others* frequently involve certain *beliefs about the reactor herself*, then many of the retributive RAs *are* grounded on the false "meta-belief" that they are merely about and merely caused by beliefs about the histories and intentions of others. Insofar as the reactor is *ignorant* of the real grounds of his or her own retributive attitudes, the reactor *lacks self-knowledge*. Insofar as *lack of self-knowledge* is a defect, the RAs sustained by such a lack are themselves defective in that they are grounded in false beliefs about the real grounds of one's own RAs. Therefore, since the genesis of so many other-directed RAs involves self deception or, at the least, a failure of self-knowledge, Strawson's reconciliationist project of demonstrating that the RAs are generally free of questionable facts *is* deeply problematical.

Moreover, we shall see that concerns about lack of self-knowledge, epistemic rationality, and dubious "facts" are not the only problems raised by the "seamy side" of the retributive RAs. Upon consideration of the seamy side's more modest versions, deep skeptical concerns arise about the *human capacity* of attaining the *impartial or moral point of*

view. Strawson argues that it *is* the "impersonal" or "dis-interested" character of retributive emotions that entitle them to the qualification "moral" (14). If, however, it turns out that many retributive attitudes adopted towards *others* presuppose beliefs and attitudes the *reactor* adopts towards *herself*, e.g., reflexive beliefs involving *her* ethnic origins, *her* familial attachments, *her* own role as a victim of a similar sort and so on, then it is hard to see how the retributive RAs — *as* they occur in the natural world — *are* entitled to the qualification "moral" by Strawson's own criteria. Contrary to what Strawson might think, the "seamy side" of the RAs *does* present problems for reconciling the disputants in the moral responsibility debate — and the dark side need *not* even be typical to *taint the RAs*. Tainted goods tend to make one lose one's appetite. Tainted RAs will make the libertarian incompatibilist lose the appetite for reconciliation.

The idea that the RAs have an important place in the traditional dispute over the problem of the compatibility of moral responsibility and determinism originates with P.F. Strawson's "Freedom and Resentment". Any investigation of the general rationality of the retributive RAs must take this seminal article as its point of departure. Other important contributions to this inquiry include: on the compatibilist side, Jonathan Bennett's "Accountability", R. Jay Wallace's *Responsibility and the Moral Sentiments* and Gary Watson's "Responsibility and the Limits of Evil," and such incompatibilists as Galen Strawson's *Freedom and Belief* and Derek Pereboom's "Determinism *al Dente*".

It seems to me that the contemporary debate between Strawson and his incompatibilist opponents over the theoretical rationality of the RAs have neglected to adequately consider the nature, pervasiveness, and dialectical upshot of the "seamy side" of the retributive attitudes. This neglected but crucial aspect of the reactive stance includes, or

so I shall argue, dubious epistemic and moral beliefs involving a lack of self-knowledge and partiality, if not pathology and sadism. We will see the explanatory need to posit devices of projection and defense against psychic pain, *and* more *modest* versions of seaminess like parochialism, and personal historical beliefs. But, the contemporary dialectic between optimists (P.F. Strawson, Jonathan Bennett, Gary Watson and R. Jay Wallace) about the compatibility of moral responsibility and determinism, on the one hand, and incompatibilist pessimists (Galen Strawson and Derk Pereboom) on the other, share an almost single-minded focus on the subject's beliefs about the *object* of the retributive RAs, e.g., beliefs about *another's* intentions, childhood history, and so on. What the contemporary discussion, all but neglects is any thorough consideration of the explanatory roles which the *subject's* beliefs and attitudes *towards him or herself* might play in generating retributive RAs towards others. In this thesis, I intend to fill this gap in the recent literature.

What is the incompatibilist response to Strawson and the other compatibilists? In his important book, *Freedom and Belief*, and in several essays, Galen Strawson, an incompatibilist determinist, insists, among other things, that Strawson senior's intuitions and anthropological observations are seriously flawed. Galen argues over and above the "impossibility of self-determination" required by any satisfactory conception of moral responsibility (Galen Strawson, The Impossibility…) changes in one's metaphysical beliefs, e.g. about agent-causes or noumenal selves *do*, as a matter of fact, play a crucial role in engendering moral responses. Galen maintains the truth of determinism itself does not make it "rational to try to adopt the objective attitude", and extirpate the reactive attitudes from our lives ("On Freedom and …" 70, 74). He concludes however, that we ought to be skeptical about P.F. Strawson's attempt to vindicate the theoretical rationality of the RAs because in

assuming that only minimal beliefs about intentions, excuses and exemptions ground the RAs, he begs the question against the pessimist. Galen insists that abandoning one's RAs *is* as natural a response to the truth of determinism as retaining them.

Rejecting P.F. Strawson's attempt to vindicate the RAs, other incompatibilists affirm Galen Strawson's conclusion that the RAs *do* respond to beliefs about determinism. Further attempts to articulate the metaphysically problematic beliefs include Derk Pereboom's "Determinism *al Dente*". Pereboom agrees with Galen Strawson arguing the RAs come with heavy metaphysical presuppositions and P.F. Strawson is mistaken to worry that the loss of the RAs render inter-personal relationships unsatisfying. According to Pereboom, emotional surrogates compatible with the rejection of true moral desert would survive a belief in determinism, and save our relationships from being diminished.

While incompatibilists differ about the content of the theoretical beliefs they think the RAs presuppose, all seem *united* against the Strawsonian compatibilist anthropological observation that our moral responses to others *do not* go beyond the Strawsonian minimum. Incompatibilists insist that when we carefully observe our moral practices and reactions, we *do* see their involvement with beliefs about determinism. The important dialectical point is this: the contemporary dispute between the optimists and the pessimists about moral responsibility has, it seems, reached an impasse on a matter of *philosophical anthropology*, namely whether the retributive RAs presuppose robust metaphysical beliefs. On the one hand, Strawsonian compatibilists maintain our moral reactions to others *do not* typically come to exist or respond to information about determinism. On the other hand, incompatibilists insist that our moral reactions *do* involve theoretical beliefs about determinism. In this thesis I aim to ease, if not break, the current impasse.

How so? I think the incompatibilists are right — at least to the extent they insist our moral reactions *do* involve *dubious* beliefs. My own "philosophical anthropology" is limited however to the narrower, but nevertheless controversial claim that *the nature of the dubious beliefs* presupposed by many, if not all, retributive RAs are of the *psychological* — not necessarily the metaphysical — sort. Since the dubious psychological beliefs I attempt to identity *do not* entail any particular metaphysical beliefs about determinism *per se*, I *cannot* claim to have resolved the perennial dispute about the compatibility of moral responsibility with metaphysical determinism. However, I do launch a version of pessimism (if not incompatibilism) in the hopes of reinforcing, if nothing else, the pessimist's "justified sense" that the RAs *do* include deeply problematic theoretical assumptions. For my purposes here I focus on the dubious psychological beliefs involved with the "seamy" side of the RAs and its dialectical implications.

How seamy *is* the seamy side of our reactive attitudes? A very *strong* version of my hypothesis would hold that *all* tokens of the moral RAs presuppose at least some unacknowledged reflexive beliefs about the subject's own history and current attitudes. The *weaker*, perhaps more plausible, version is that such unacknowledged reflexive beliefs underlie the RAs *much more pervasively* in human life than Strawsonian "minimalist" philosophical anthropology suggests. Given my general suspicions about the global unhealthiness of the RAs, I would love to be able to establish the strong thesis. However, this task is beyond me, and I am prepared to grant that some, perhaps many, instances of the RAs *are* appropriate. But, is the *weaker version* of my thesis worth defending? Again, it is worth repeating that Strawson makes much of the claim that it "is an exaggerated horror, itself suspect, which would make us unable to acknowledge the facts (about the minimal

presuppositions of the RAs) because of the seamy side of the facts" (25). Strawson continues maintaining that:

> ...psychological studies have made us rightly mistrustful of many particular manifestations of the attitudes I have spoken of. They are a prime realm of self-deception, of the ambiguous and the shady, of guilt transference, unconscious sadism and the rest. But it is an exaggerated horror itself suspect, which would make us unable to acknowledge the facts because of the seamy side of the facts (25).

The *weaker version* of my thesis *is* worth defending, that the existence of the "seamy side" ought *not* be taken as lightly as Strawson maintains — even if some RA are appropriate — since the "facts" involved in the "seamy side" of retributive resentment and indignation involve pathological forms (as Strawson acknowledges above), *and* more modest forms he seems to neglect. When we consider the explanatory power of *a more complete spectrum* of "seamy" retributive attitudes, for instance, beliefs and attitudes involving prior victim hood, ethnicity, lust and so on, then the pervasiveness of the "seamy side of the facts" should give the incompatibilist significant pause. Seamy retributive attitudes involving pathology, reflexive moral beliefs, dubious desires, personal prejudice, and so on, also involve at worst *self-deception* and, at best, a deep *lack of self-knowledge and partiality* on the part of the reactor. If this, a more *robust, seamier alternative* to Strawsonian minimalism, explains as many of our anthropological observations about inter-personal RAs as I think, this will indicate that philosophers tend to *underestimate* the dangers and dialectical implications of the reactive stance, *not* exaggerate them, as Strawson suggests.

My "Senecan"/"Nietzchean" challenge to Strawson's "Aristotelian/Humean" vindication of the reactive emotions seems to depend on the truth of an *empirical* theory about the genesis of these emotions. Is this a problem for a *philosophical* argument? I do

not think so. My aim is modest and does *not* depend on the universal applicability of the projection hypothesis. Inspired by the explanatory turn in recent literature, I do argue however, that positing such seamy beliefs and attitudes as reflexive guilt, partiality, prejudice, proximity and so on, provide, at least frequently, a better explanation for reactive experience than Strawsonian minimalist presuppositions — including impartial sympathy. Moreover, *it is* interesting enough that, in spite of Strawson's assurances to the contrary, there *are facts* about the RAs that Strawsonian belief conditions *do not* explain. And, it is also important to show that Strawson is mistaken to think that "being true to the facts as we know them," *is sufficient* to save the RAs from frequent charges of theoretical irrationality. Bringing the reflexive *subject* back into the moral responsibility debate reveals that the thesis of determinism is *not* the only condition upon which optimists and pessimists ought to be concerned. Since the genesis of so many other-directed RAs involves self deception, or at the least, a failure of self-knowledge, Strawson's reconciliationist project of demonstrating that the RAs are generally free of questionable facts *is* deeply problematical.

In **Chapter One, Section One**, I outline an important part of the stand off between "optimists" about moral responsibility, and, their metaphysically attended "pessimist" rivals. I discuss Strawson's hope that his methodological and conceptual toolbox, including what he maintains is the "*vital thing*" missing from the optimists account, is sufficiently attractive to bring the parties together — *without* metaphysical accompaniment.

In **Section Two of Chapter One**, I sketch the main incompatibilist responses to Strawson's proposal for unification since the appearance of "Freedom and Resentment." I point out the current impasse seems to involve disputes about *matters of fact*, namely, the nature of the descriptive and normative beliefs, presupposed by the reactive attitudes.

In **Chapter Two, Section One**, I discuss loosening the post-Strawsonian log jamb over the belief presuppositions of the RAs by, in part, articulating latent methodological themes in this, the post-Strawsonian wave. I articulate this "new" mode of exploring questions about moral responsibility, suggesting a shift to explanatory potency and inference to the best explanation, from the current over reliance on conceptual analysis and philosophical intuitions. I make more explicit the *explanatory turn* we begin to sense in Gary Watson's important essay, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme". We explore Watson's argument that we ought to be deeply skeptical about the theoretical rationality of the RAs since philosophical anthropology does not, as Strawson maintains, reveal the genealogy of the RAs to be solely non-controversial beliefs. I argue in response that Watson is mistaken by suggesting that Strawson cannot revise his account with a historical condition — without conceding to the incompatibilist. However bleak the prospects, articulating a satisfactory historical condition remains open to compatibilists. Providing they argue against incompatibilist generalization arguments, compatibilists it seems, can be historicists.

In **Chapter Two, Section Two** we discuss one such historicist compatibilist, R. Jay Wallace, whose *Responsibility and the Moral Sentiments* also involves a latent methodological shift to explanatory potency. Wallace echoes Watson's worry about the explanatory inadequacies of Strawsonian minimalism. He argues the "sensitivity" of retributive reactions are *not* explicable *merely* in terms of beliefs about whether an agent's intentions and histories exclude him from moral and inter-personal communities. As we shall see, Wallace suggests moral responsibility is, what Fisher and Ravizza call "genuinely historical", and not as Strawson implies, "epistemically historical" (Fischer and Ravizza 173-

194). Wallace suggests moral responsibility cannot, in principle, be explained *solely* in terms of the present psychological states of agents, for moral responsibility requires abilities to grasp and apply moral reasons, including the "fundamental commitment to fairness", that are themselves acquired *from* historical sources (Wallace 102). Rather than beliefs about an agent's abused childhood as a mere *means* to some current feature about them, Wallace argues historical deprivation may show one's capacity for rational self-control has been damaged, and acting according to moral reasons is, in varying degrees, difficult.

Although, as we will see, Wallace, more than anyone else, stresses the importance responsiveness to historical considerations plays for the *fairness* of holding responsible, and seems deeply troubled by the role dubious reflexive beliefs play in the retributive attitudes, I will argue he *fails* to articulate among the most important aspects of the "seamy side" which jeopardizes the whole Strawsonian reconciliation strategy, namely, emotional *akrasia.* Phenomena such as *akrasia* dull the luster of what Strawson hopes is the "glittering prize" of the retributive attitudes. Wallace's treatment of *akrasia* taints the RAs because his account makes moral *judgments the* central feature in holding responsible, *not* the reactive attitudes themselves. Like Strawson and Watson, even Wallace fails, in the end, to take the "seamy side" seriously enough.

Although they notice certain explanatory inadequacies of Strawsonian minimalism, Watson and Wallace it seems, fail to provide him with the *dialectical leverage* needed to vindicate the retributive attitudes. Perhaps, hypotheses for important retributive experience centering upon the reactor's beliefs and attitudes *about others*, themselves, involve important explanatory gaps. The latent emphasis on explanatory potency has however, provided

inspiration for my own hypothesis involving the content of the reactor's beliefs *about him or her self.*

In **Chapter Three, Section One**, I discuss *my* take on the explanatory inadequacies of Strawsonian minimalism and present *my* hypothesis involving dubious reflexive beliefs and attitudes. I offer empirical data supporting the notions that many, if not all, RAs presuppose what amount to *dubious epistemic and moral beliefs and attitudes* about *our own* history and interests. I argue that *over and above* any beliefs about others, on many occasions important aspects of our RAs are *best explained* in terms of the "seamy side" of the retributive RAs involving certain unacknowledged beliefs and attitudes the reactor has towards him or herself.

I continue the argument in **Chapter Three, Section Two**, maintaining that *one way* the contingent reflexive beliefs of the self influence the duration, intensity, responsiveness — even the existence — of the retributive attitudes is the *projection of guilt*. I argue this neurotic (if not pathological) version of the seamy side is problematic because the reactor believes his or her retributive attitudes are *solely about others*. But, to the extent the retributive attitudes involve the projection of guilt, they are *not solely about others*. Therefore, the retributive attitudes that involve the projection of guilt involve false beliefs.

I *extend* this argument to **Chapter Three, Sections Three and Four**, that over and above neurosis and pathology, the "seamy side" of the retributive attitudes also involves more *modest versions*. Reflexive beliefs and attitudes involving one's proximity to the perpetrator, the person harmed, shared history, ethnicity, religion, victim hood and so on, frequently provide a *better explanation* than Strawsonian minimalism alone for retributive experience. I argue that although such "modest" retributive attitudes appear, on the face of it,

less worrisome than their pathological counterparts, they *are* "seamy" nonetheless *for* both presuppose dubious moral and epistemic beliefs involving lapses of impartiality, if not rationality.

In **Chapter Four**, I examine two Strawsonian rejoinders to my argument that he seriously underestimates the nature, extent, and *dialectical import* of the RA's "seamy side". In **Section One**, I consider a Strawsonian defense of the retributive attitudes that appeals to a strong role for impartiality ensuring *institutions*. Although the sober second thought of institutional judgments may go some distance to ensure against retributive bias, an increased need for impersonality, rules and regulations will leave the libertarian more suspicious about the role reactive feelings play within the actual practice of moral responsibility. I argue however, while institutional safeguards sooth some suspicions, they *stunt* Strawson's reconciliation strategy. In **Section Two**, I consider a further Strawsonian attempts to calm worries about the epistemic and moral values of retributive RAs, by appealing to the importance of *impartial sympathy* and the care and concern for our *common humanity*. I argue that any such defense *assumes too much*, since, for instance, deep feelings of care and sympathy for human beings *do* seem compatible with adopting the objective stance.

Finally, in **Chapter Five**, I close with a few reflections on what *this* more pervasive but modest version of the "seamy side" of reactivity might mean for Strawson's reconciliation project. In short, *worries about modest but seamy partiality* of the retributive attitudes seem to *taint* the RAs for Strawson's reconciliationist hopes; the pessimist becomes disinterested, for any increased need for impersonality, rules and regulations will leave the libertarian uninspired and lacking the vitality he misses under consequentialism. More suspicious than ever about the importance reactive feelings play within the actual practice of

moral responsibility, the pessimist becomes *skeptical* about the moral and epistemic prospects of a naturalistic account of moral responsibility.

I have *four goals* in this thesis:

First, I hope to show that the kind of philosophical anthropology Strawson employs in his theoretical vindication of the RAs is far *too limited in scope* to achieve his central goal of reconciling compatibilists and incompatibilists.

Second, I *broaden the range* of evidence that characterizes our anthropological observations about the nature and existence of our other-directed RAs and enter the best explanation of why we often evince them. The range of our other directed RAs, most notably, their Duration, Intensity, Responsiveness to historical pleas and Existence (the D.I.R.E. aspects of reactive attitudes) suggests that their best explanation presupposes that their subject possesses certain *moral and psychological beliefs about him or her self.* For instance, where S resents R for doing *p* to her, then S believes that *she is also* guilty for committing *p* — or at least some thematically related action of type *q*.

Third, on a more methodological level, I hope to take the current dispute over the presuppositions of the RAs, which divides Strawson and the pessimists beyond mere conceptual intuitions and dubious observations about philosophical anthropology. In short, I want to add an *explanatory dimension* to the debate.

Fourth, I hope to show that the self-reflexive beliefs frequently, if not globally, presupposed by the subject's RAs, are philosophically and psychologically dubious in a way which further undermines Strawson's attempt to save them from incompatibilist doubts, and thus to achieve a détente between incompatibilists and compatibilists. Even *if* Strawson is

correct to suppose that beliefs about the truth of metaphysical indeterminism are *not* presupposed by RAs, this is *not enough* to protect his theory of moral responsibility from criticism: the reactive attitudes do involve *other false beliefs* about the facts, e.g., self and other, and *dubious psychological beliefs involving partiality*, if not *pathology*.

Contrary to what Strawson might think, the "seamy side" of the RAs need *not* be "exaggerated" to give us reason to be worried about their rationality, partiality, and importance to human flourishing.

Notes To Introduction

1. My line of investigation in this thesis has a long history of the relationship between reason and emotion in both ancient and early modern philosophy. It is illuminating to view P.F. Strawson as providing a contemporary criticism of the ancient Stoical understanding of moral emotions or reactive attitudes. In *De Ira*, Seneca argued that all emotions, including moral emotions, are irrational because they presuppose false beliefs about human nature. Accordingly, Seneca also argued that getting the facts straight about human beings, especially about what really motivates human action, would, or ought to, dispel moral emotions like resentment and indignation.

2. I am indebted to Professor David Zimmerman for helping me understand the importance of a satisfactory account of autonomy, including the source of our desires, for any satisfactory theory of moral responsibility.

3. P.F. Strawson, "Freedom and Resentment," 1-25. Unless noted otherwise, all citations will be from this work.

4. In this thesis I employ, roughly, R. Jay Wallace's amended version (25-33) of Strawson's moral reactive attitudes, according to which the moral reactive attitudes involved with the practice of moral responsibility are the *other* directed negative attitudes of resentment and indignation, *and* the self directed negative attitude of guilt. Unless noted otherwise however, when I use the terms "retributive RAs" or "retributive reactive attitudes", I am speaking more specifically about the *other*

directed negative moral attitudes of resentment and indignation. I will call the third and final moral reactive attitude of guilt by name.

5. In *The Nichomachean Ethics* (Aristotle Vol. IX), Aristotle foreshadows the Strawsonian response, arguing against Seneca that our moral emotions do not, at least globally, violate standards of rationality. For instance, Aristotle argues the moral emotions do not generally violate *practical* standards of good reasoning, but rather, contribute greatly to human flourishing. Strawson embraces the Aristotelian thesis that there is no global, practical reason to give up the reactive emotions. In this thesis I focus on Strawson's reconciliationist hopes. But, persuading the pessimist libertarian to *trade* her prized metaphysical commitments for the reactive attitudes *does* require the assurance the attitudes *are* free of controversial or false beliefs-theoretical or otherwise. We will see, although there is no *global* theoretical nor practical reason to *completely* eradicate the retributive RAs from our lives, deep worries about the pervasiveness of theoretical and practical irrationality remain. Contrary to what Strawson might think, worries about the retributive RAs are *not* "exaggerated" (25).

In the early modern period, David Hume took up a version of the Aristotelian defense of the reactive emotions. Hume famously held that our passions are, strictly speaking, beyond the assessment of theoretical rationality. ("Reason is and ought only to be the slave of the passions") (Hume, *A Treatise Of*, Book II, Part III, 415). Nonetheless, he was a kind of "cognitivist" about the intentional structure of passions, or reactive attitudes like pride suggesting that at least some passions are, at least in

part, constituted by beliefs. So, like both Seneca and Aristotle, Hume *did* seem to allow that the passions or moral reactive attitudes *can* be rationally assessed or based on empirical beliefs about their objects. However, in contrast to Seneca, Hume maintained that moral emotions do not *always* presuppose false, pathological, or otherwise dubious beliefs. According to Hume, moral emotions like sympathy and a sense of justice, though subject to certain conditions of appropriateness, are natural and desirable, even indispensable features of a civilized human life.

6. In the modern period Friedrich Nietzsche and, to some extent, Sigmund Freud resurrected Senecan skepticism about the philosophical integrity of moral responses to others, like resentment and indignation. Like Seneca, Nietzsche, in *The Genealogy of Morals*, argues that all retributive emotions are pathological. But, rather than embracing the Senecan notion that *all* persons — except the Stoic sage — are sick (to one degree or another), and thus prone to anger and resentment towards others as a means of distraction, Nietzsche argues that *only* the weak are sick and suffer from this tendency. According to Nietzsche, the weak develop Judeo-Christian or, "slave morality", because they need the reactive attitudes of resentment and indignation to deflect themselves from the painful realization of their own frailties. The weak resent the strong, trying to make the strong feel guilty, in order to bring them down, and boost themselves up. In some respects Nietzsche shares with Seneca a premonition of what we will see is Freud's theory of *projection.*

My sympathies are more akin to Senecan egalitarianism then to Nietzschean elitism and I reject most of what he says about the nobility of the "masters" and the

repugnancy of the "slaves". In somewhat of a Nietzschean spirit however, I argue that the "reasonableness" of a person's reactive attitudes towards others is suspect, since, unbeknownst to the person experiencing the attitude, they frequently pre-suppose self-referential beliefs about the person's own inadequacy, personal preferences and so on. Nietzsche seems to foreshadow the central feature of my argument against Strawson since like Nietzsche I argue that the theoretical rationality of the reactive attitudes is seriously suspect because their *source* or genealogy includes (at least frequently) dubious beliefs about their subject.

CHAPTER ONE

Strawson's Reconciliationist Project

P.F. Strawson argues that *being* morally responsible is grounded in the practice of *appropriately holding* people morally responsible, that the vital core of this practice is our natural proneness to take up the *reactive stance* towards people, and that the general appropriateness of our feeling resentment, indignation and the like is not jeopardized by the "facts as we know them". He concludes that this is enough to vindicate the very existence of moral responsibility in the face of incompatibilist worries. In Section One of this Chapter, I outline the basic elements in the methodological and conceptual toolbox, which Strawson puts to work in his attempt to reconcile the optimist with the pessimist about moral responsibility. In Section Two, I sketch the main incompatibilist responses to Strawson from several sources after "Freedom and Resentment" was first published.

Section One: Strawsonian Minimalism and the Vindication of the Reactive Attitudes

Strawson's account of moral responsibility is designed to resolve the perennial feud between the compatibilist and incompatibilist. Concerning moral responsibility the incompatibilist ("the pessimists") worries that if determinism is true, then human actions and their origins are caused in ways that are not compatible with moral responsibility. Incompatibilists are divided into two camps, "hard determinists" who believe that the world is actually determined and that moral responsibility is illusory, and "libertarians" who believe that there is moral responsibility in the world since it is indeterministic, and that such a metaphysical structure allows room for noumenal selves, contra-causal freedom, agent-

causation, or some such freedom permitting thing. Compatibilists ("the optimists"), on the other hand, believe that supposing the truth of metaphysical determinism gives us no reason to worry about the philosophical and practical integrity of the institutions, concepts and practices associated with moral responsibility. According to the compatibilist, the kind of free will presupposed by praising and blaming, rewarding and punishing fits nicely with the existence of causal explanations for human actions.

Strawson's goal is to reconcile the pessimists and the optimists by, as he puts it, exacting "a formal withdrawal on the one side in return for a substantial concession on the other" (2). The reactive attitude centered compatibilist account of moral responsibility admits that the kind of exclusively consequentialist rationale offered by traditional optimists from Hume to Schlick, "leaves out something vital" from what we are doing when we hold people responsible. He grants, moreover, that "the pessimist is rightly anxious to get this vital thing back." However, he worries that "in the grips of his anxiety (the pessimist) feels that he has to go beyond the facts as we know them," and that he can secure that "vital thing" only if "there is the further fact that determinism is false" (2).

Strawson argues that the libertarian way of preserving the "vital thing" (*vis-a-vis* the practice of responsibility) is philosophically dubious or excessively controversial. According to Strawson, the pessimist need not opt for anything as "panicky" (25) as an indeterminist metaphysical self in order to restore that "vital thing" missing from the compatibilist conception of responsibility *qua* expediency. Rather, the two sides can be reconciled, provided that is, that the consequentialist compatibilist fulfills his side of the bargain, and admits that there is more to the practice of holding people responsible than mere behavior control, and the incompatibilist abandons metaphysical presuppositions involving

controversial notions of the self and human action. But, what is this "something more," this "vital thing," which the optimist can concede to the pessimist in exchange for the latter's giving up his baroque metaphysical notions? Strawson argues that the "vital thing" that the pessimist thinks is missing from the consequentialist conception of responsibility must include "desert, responsibility, guilt, condemnation, and justice". Rather than responsibility *qua* utility, the pessimist longs for responsibility *qua* "*all* we mean, when speaking the language of morals" (23). While the "vital thing" missing from the consequentialist conception of condemnation and punishment must be both compatibilist friendly *and* vindicate such important moral notions as "desert" and "justice", Strawson makes it clear that whatever this "vital thing" is, it must *not* go "beyond the facts as we know them."

Strawson shifts the focus of the free will/determinism debate by emphasizing the kinds of moral *emotions* presupposed by the practice of responsibility — including, namely, the "non-detached attitudes and reactions of people directly involved in transactions with others" (4), which are grounded in "the very great importance that we attach to the attitudes and intentions towards us of other human beings" (4). Strawson calls these "reactive attitudes" (6). They include emotions like resentment, indignation, gratitude, and forgiveness (5). They are "personal feelings and reactions (which) depend upon, or involve, our beliefs about these attitudes and intentions (of others)" (5).

Strawson contrasts the reactive stance associated with ordinary inter-personal relationships with what he calls the "objective" stance. The pivotal distinction between the two stances is that abandoning the reactive stance in favor of the objective stance is to see the other as an object of curiosity, care or manipulation, rather than as one to whom we are involved in certain "adult" reciprocal relationships. The objective stance is frequently

"teleological" or goal-directed (Bennett 36) and, although it *is* compatible with a robust array of emotions, e.g., repulsion, fear, pity and some forms of love, Strawson maintains that taking up the objective stance is *not* compatible with the RAs that characterize our "attachments" to persons, such as resentment, gratitude, indignation, forgiveness, or "the sort of love that two adults can sometimes be said to feel reciprocally for each other" (9).

Strawson argues we are *not* the kind of creatures that permanently reside in the realm of objectivity. Rather, we are *prone* to adopt the reactive stance towards others and ourselves *because* of "human nature" (16). What *is* human nature according to Strawson? Strawson maintains that humans are, at least for the most part, deeply concerned about the good and ill will of people towards *not* just ourselves — but also *others*. More specifically, Strawson claims the retributive RAs are motivated by impartial ("...vicarious or impersonal or disinterested..." (14)) care and concern for the well being of "our common humanity". They express the "generalized" sympathy one feels "on behalf of another" when our own "interest and dignity are not involved" (14). We are *not*, according to Strawson, "moral solipsists" (16), or "egoists" (18). Strawson hopes to persuade the pessimist that the reactive stance is an integral part of what makes us distinctly human and that we do not, for the most part, have "a choice" about its adoption (12). Strawson maintains he is not so much concerned with why humans are prone to reactivity (6), but more with articulating the variations or patterned ways humans do, in fact, abandon the reactive for the objective stance. He is convinced that this sort of descriptive investigation, a kind of "philosophical anthropology," will also reveal the conditions under which we *ought* to abandon the reactive for the objective stance.

Thus, the first step in Strawson's reconciliationist strategy is basically a descriptive one: to argue that *as a matter of fact* our RAs are not, as the pessimist maintains, based on

controversial philosophical beliefs about agent causes, noumenal selves or any other troubling notions involving the falsity of determinism. Rather than dubious philosophical beliefs, the practice of moral responsibility and the RAs pre-suppose nothing more controversial than beliefs about the intentions and histories of others. Strawson argues the world of personal relationships reveals the RAs respond, at least typically, to our concern that *good will* be shown towards all humanity. The basic demand for good will is grounded in our care and concern that, in general, all humanity be shown a certain degree of good will, or at least, the absence of "active ill will or indifferent disregard" (14). We react in various ways to the quality of another's intentions as displayed in their attitudes and actions. More specifically, Strawson argues, we typically adopt negative reactive stances when we believe that the other person fulfills two minimal psychological conditions: first, that he has failed to satisfy the "basic demand," and second, that he cannot plead any of the standard excusing or exempting conditions.

For instance, when we come to believe that another person has harmed us malevolently, manipulatively, or as the result of avoidable ignorance, we are inclined to feel resentment. We become indignant upon realizing that a person has exhibited a similar absence of good will towards a third party. We adopt the reactive stance *provided* that is, that we also believe the offender is not subject to any excusing or exempting conditions.

The qualification just noted is significant, however, for under certain conditions we do inhibit our RAs, sometimes even when we do believe that the other person has failed to meet the basic demand. Strawson notes that there are two kinds of inhibiting conditions: The first involves the standard list of *excuses*, for example, the agent is ignorant or coerced. When we excuse another, we do believe that the person is "a responsible agent, but we see

the injury as one for which he was not fully, or at all, responsible" (7). For instance, committing harm inadvertently or from ignorance is compatible with an agent's upholding the basic demand.

When the second kind of inhibiting condition, an exemption, obtains, then the other person is not subject to the basic demand in the first place (8). By virtue of being psychologically or morally underdeveloped or disabled exempted agents lack "membership in the moral community." Therefore, they are not subject to the demand for good will. Examples are children, the insane, and those who are, in general, incapable of controlling themselves in accordance with the basic demand. Also, those who are typically capable of controlling themselves in accordance with the basic demand, but fail to do so on a particular occasion by virtue of being drugged, hypnotized, or subject to great stress.

According to Strawsonian anthropological philosophy, RAs like resentment and indignation *do* have conditions under which we believe they are *appropriate*. When RAs do not conform to these conditions, they become subject to criticism. Like most emotions, RAs are *intentional* states, which have proposition-like objects. Therefore, RAs are, at least in part, constituted by *beliefs*. Since RAs are intentional, RAs have conditions of *theoretical* appropriateness. In other words, one condition of their appropriateness or inappropriateness lies in certain features of the empirical beliefs of the person who evinces them, namely, whether the beliefs are true, justifiable and consistent. So, for instance, A appropriately resents B *only if* A's belief that B has violated the basic demand for good will is itself justifiable, consistent with the facts and true, and B has, as a matter of empirical fact, violated the basic demand for good will.

For Strawson, these conditions of the theoretical appropriateness or epistemic rationality of the RA are *minimal* or unproblematic.[1] They do not, according to him, involve controversial metaphysical beliefs about undetermined selves, involving only commonplace empirical beliefs about whether the other has failed to satisfy the basic demand for good will, and whether the other is subject to the standard excusing and exempting conditions. Strawson argues that, because these appropriateness conditions presuppose neither dubious metaphysical beliefs involving the falsity of determinism nor any controversial notion of self, they are immune from incompatibilist theoretical worries.

Strawson asks the pessimist to consider this crucial question:

> *Would* or *should* the acceptance of the truth of the thesis of determinism lead to the decay or the repudiation of all such attitudes? Would it mean the end to gratitude, resentment, and forgiveness; of all reciprocated adult loves; of all the essentially *personal* antagonisms (10, italics mine).

Strawson's answer to *both* the anthropological ("would") and the normative ("should") question is a resounding *no*. Because the RAs depend only upon minimal beliefs about an agent's abnormality or immaturity, and since "it cannot be a consequence of determinism" that human abnormality or immaturity is "the universal condition" (11), he insists that "no acceptance of the truth of the thesis of determinism" *would* lead to our repudiation of the RAs. He goes even further, insisting that *even if* the reactive stance did somehow theoretically presuppose the falsity of determinism, our "human commitment to participation in ordinary inter-personal relationships" is too deep-rooted to allow any "general theoretical conviction" to end relationships as we know them. According to Strawson, the abandonment of the reactive stance entails that we relegate ourselves to the "human isolation" of the objective stance. Strawson insists, that no "general theoretical

truths" could provide enough *practical* reason for our succumbing to the purely objective condition which is barely recognizable as human (11-12).[2]

Note that Strawson's reconciliation project seeks to persuade the pessimist that there is neither *theoretical* nor *practical* reason to abandon the retributive RAs and that it is *these attitudes* upon which any satisfactory practice of moral responsibility depends. The consequentialist conception of moral responsibility leaves out the retributive or the "moral" RAs (14), and Strawson argues that it is *their absence* that leaves the pessimist with the "justified sense" that there *is* some "vital thing" missing from the consequentialist conception. In his effort then to assuage pessimist worries, Strawson argues for two distinct theses: 1) that the retributive RAs include no general metaphysical beliefs among the beliefs they do presuppose and 2) that, even if they did, there would be no general *practical* reason for us to give them up.

Section Two: Incompatibilist Objections, Dialectical Impasse

Does Strawson's reconciliation of the optimist and the pessimist work? Does he succeed in persuading the pessimist that the RAs presuppose only straight forward empirical "facts as we know them," and not the controversial metaphysics of the libertarian self? It seems not. Why should the pessimist believe that *the* typical anthropological landscape of our inter-personal relationships and moral reactions is, as Strawson maintains, *the* philosophically appropriate landscape? What reasons has Strawson given for believing that our moral reactions in fact do respond, and moreover ought to respond, *only* to the minimal belief and attitude conditions he articulates? For instance, he may be right that one's beliefs about determinism, as a matter of anthropological fact, make no empirical difference to our

feelings of resentment and indignation. And he may also be right that we do, at least typically, respond to information about the psychological states of others, e.g., the nature of their intentions and whether they satisfy standard excusing and exempting conditions. But, the truth of these descriptive claims seems insufficient to satisfy the pessimist. After all, the pessimist will respond to Strawson by arguing that the truth of the determinist thesis itself provides a *general* excusing or exempting condition, precisely because it undermines the sort of human freedom necessary to sustain moral responsibility.

In *Freedom and Belief*, and in several essays, Galen Strawson, an incompatibilist determinist, launches this version of the pessimist response (among others). Over and beyond the problems the "impossibility of self-determination" (Galen Strawson "On Freedom and Resentment"67, and *Freedom and Belief*, ch.2, esp.2.1) raises for an intelligible moral responsibility, Galen also insists that Strawson senior's anthropological observations and conceptual intuitions are seriously flawed. According to Galen, empirical observation reveals that changes in one's metaphysical beliefs, e.g., about agent-causes or noumenal selves do as a matter of fact (*contra* P.F. Strawson) play a crucial role in altering the moral responses we adopt toward people:

> Consider a man who becomes a determinist. He is often pictured as being faced first and foremost with the problem of what he is to make of other people, given his new belief. ...It seems that most people would find abandonment of the ordinary, strong notion of responsibility intolerable, not to say practically speaking impossible, from a social point of view... For it is not as if one can excise one's inclination to praise and blame people while leaving all one's other attitudes to them untouched. If determinism is called upon to justify such excision, (reactive attitudes are) thereby put at risk... (Galen Strawson, "On Freedom and Resentment" 68).

Galen maintains that upon assuming the truth of determinism most people would *not* be inclined to adopt reactive attitudes like resentment and indignation towards one and other. Compatibilists like Strawson senior thus seriously underestimate:

> ...the equal naturalness of the pessimists' position... The fact that the basic incompatibilist intuition has such power for us is as much a natural fact about cognitive beings like ourselves as is the fact of our quite unreflective commitment to the reactive (Galen Strawson 70).

As we can see, Galen maintains that careful observations of our inter-personal relationships reveal that our moral attitudes *do* quite naturally cease or respond to metaphysical beliefs about determinism. Moreover, he argues that pessimists have different, yet equally compelling, philosophical intuitions about the necessary conditions for moral responsibility from those of the optimist. His point is that these intuitions are resolutely *in*compatibilist. Further, he suggests that Strawson senior begs the question when he *simply asserts* his conceptual intuitions, namely, that the RAs pre-suppose only minimal beliefs about intentions, excuses and exemptions, but no beliefs about determinism.

It seems, then, that the debate over the implications of the RAs for the reality of moral responsibility faces a logjam or impasse of sorts. On the one hand, Strawson senior's descriptive anthropology and conceptual analysis maintain that appropriate RAs *do not* presuppose beliefs about determinism, and are thus *beyond* specifically theoretical criticism. On the other hand, the descriptive anthropology and conceptual analysis of pessimists like Galen Strawson maintain that the sincere entertainment of these RAs *does* presuppose robust beliefs about the metaphysical structure of the world which human beings inhabit. Contrary to P.F. Strawson's complacent assumption, there appears to be a genuine philosophical standoff here.

Other incompatibilists line up alongside Galen Strawson and respond in similar fashion to P.F. Strawson. They too attempt to articulate the nature of those elusive beliefs that the moral RAs presuppose. For example, in "*Determinism al Dente*", Derk Pereboom, in attempting to revive the fortunes of "hard determinism", echoes Galen Strawson's argument that the reactive attitudes come with heavy philosophical baggage. He also responds to P.F. Strawson's worry that loss of the RAs would leave inter-personal relationships bankrupt. Pereboom argues that those reactive attitudes that P.F. Strawson would want to retain have "emotional analogues" that are not based on false suppositions. These analogues, according to Pereboom, would survive a belief in determinism, and are emotionally rich enough to preserve inter-personal relationships and facilitate human flourishing (Pereboom 269). Pereboom's concern with the rationality of the RAs is mainly practical. We will return to Pereboom and this practical theme in Chapter Four Section Two when we consider the role impartial sympathy plays in the good life of human flourishing — a Strawsonian defense. I note for now however, that Pereboom joins Galen Strawson in rejecting both P.F. Strawson's (compatibilist) anthropological observations and intuitions that the appropriateness conditions of the RA are minimal.

We have seen that the post Strawsonian dispute between optimists and pessimists about moral responsibility has reached an impasse. The Strawsonian compatibilist maintains that the retributive RAs are *not* grounded, either empirically or normatively, in beliefs about determinism. The incompatibilist pessimist maintains that they *are*.

Notes To Chapter One

1. As we have seen, Strawson argues that the RAs involve only "minimal" or non-controversial theoretical beliefs about the intentions and histories of persons, rather than controversial theoretical beliefs about persons and macroscopic determinism. For Strawson, reactive attitudes are theoretically rational or "appropriate" from the epistemological point of view if and only if the beliefs they presuppose about the intentions and histories of others are true and consistent with the "facts". If the beliefs they presuppose are false or inconsistent then the reactive attitudes are subject to criticism on epistemic grounds. For example, "S resents T" is open to criticism if either S's belief that T acted against an interest of his out of malice or negligence is false. Much of Strawson's article is devoted to persuading the incompatibilist that the reactive attitudes do not *involve* false beliefs, particularly about metaphysical determinism. Moreover, Strawson suggests that if our general theoretical presuppositions about the unconscious, self-deception, and the efficacy of the reactive attitudes turn out to be false, we may "have good reason" for modifying, and redirecting them or dropping the related practice (25). Strawson argues on practical grounds however, (see note following) that there can be no theoretical reason dictated by our desire to "be true to the facts as we know them", that would, or should, lead us to totally abandon the reactive stance.

2. Strawson argues, on practical grounds, that we are not rationally obligated to give up the reactive attitudes in light of any general theoretical considerations. Instead of theoretical truths, "if we could imagine... a choice in the matter, then we could

choose rationally only in the light of an assessment of the gains and losses to human life..." (13). Although the "dreams of some philosophers" (25) might be realized by so doing, it is practically inconceivable that any mere theoretical belief would or should permit giving the reactive attitudes up. Schlickian compatibilism also supports the notion that the rationality of moral practices is determined along practical grounds. But, according to Strawson, Schlick borrows his theoretical presuppositions from a "one eyed" utilitarianism (23), and, in so doing, fails to endorse the practical rationality of the retributive attitudes. Strawson, on the other hand, seems to borrow *his* theoretical presuppositions about practical reason from Aristotle. Implicit in Strawson is something like the Aristotelian notion that practical reason and *human flourishing* require the *proper emotional response* to the character of human intention and action. And, like Strawson, Aristotle maintains that at least sometimes, *the* proper emotional response *is retributive*. According to Aristotle, exhibiting "*moral* virtue" includes adopting the appropriate retributive emotion "at the right times, with reference to the right objects, towards the right people, with the right motive..." (Aristotle, *Nicomachean Ethics*, Book II). For both Strawson and Aristotle, appropriate indignation is adopted by the virtuous (Strawson's "civilized"), toward the right people, at the right time, on behalf of the ill fortune of another. For both, the failure to adopt the reactive stance thus means the failure of practical reason, and the failure to fully appreciate the value of human beings. The upshot of Strawson's/Aristotelian practical argument? There is no general practical reason to give up the reactive attitudes. To extirpate the retributive emotions from human life would leave life impoverished and lacking the very thing that makes us human, the

tendency to make moral demands upon each other: "What is wrong is to forget that these practices, and their reception, the reactions to them, really are expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes. Our practices do not merely exploit our human natures, they express them. Indeed our very understanding of the kind of efficacy these expressions of our attitudes have in turns on our remembering this" (25). For Strawson's Aristotelianism, the retributive attitudes and the network of reactive attitudes as a whole are beyond rational assessment in the sense that *who we are* is beyond rational assessment. The elimination of the retributive attitudes and the sole domination of the objective attitude thus offends "humanity" since what it is to be human just *is* to be the type of beings that react to ourselves, and others with resentment, indignation, guilt, gratitude. In short, to imagine ourselves without the reactive attitudes *is not* to imagine ourselves at all, but rather "to change the subject of philosophical anthropology" (Bilgrami 217). As we shall see in Chapter Four Section Two, it is doubtful however, that Strawson's reliance on an Aristotelian conception of practical flourishing, including an Aristotelian conception of *sympathetic selves*, provides a *non-controversial* alternative to the metaphysics of *libertarian selves*.

CHAPTER TWO

Some Explanatory Failures of Strawsonian Minimalism

We closed Chapter One noting the post-Strawsonian dispute between optimists and pessimists about moral responsibility had reached an impasse. The Strawsonian compatibilist maintains that the retributive RAs are *not* grounded, either empirically or normatively, in beliefs about determinism. The incompatibilist pessimist maintains that that they *are*. One methodological handicap the current dispute faces then is too much emphasis on controversial conceptual intuitions about the beliefs or intentional content of the RAs, and about the nature of the empirical data which "philosophical anthropology" yields. Perhaps offering a *new mode* for exploring this contentious question might help.

It seems to me that investigating the belief-presuppositions of the RA should, in addition to intuition and observation, appeal to considerations of *power*. After all, all parties in the responsibility debate want to understand *what beliefs the RAs presuppose*. I think that this common goal is better served by the search for the *best explanation* of certain features of the RAs, rather than merely by engaging in highly contentious conceptual analysis and philosophical anthropology. This shift of methodological focus forms the background of this chapter and the next (Chapter Three).

However, I should acknowledge that this shift of focus is not entirely original with me. It is *implicit* in both Gary Watson's "Responsibility and the Limits of Evil" and R. Jay Wallace's *Responsibility and the Moral Sentiments*. Both investigate how to identify the content of the beliefs, which our RAs presuppose. Each seems to employ, without explicitly saying as much, something like the method of *inference to the best explanation*. In this

chapter I try to make Watson's and Wallace's explanatory turn more explicit, by outlining what I take to be their discoveries about the RAs. First, however, I will review some of the difficulties they think inherent in Strawson's minimalist belief thesis has in explaining certain features of our RAs. Watson can't seem to find sufficient cognitive and affective resources within Strawson to accommodate important experiential aspects of the retributive RAs. He concludes that something over and above Strawsonian minimalism *is* required to accommodate the nature of our reactions to (what are purportedly) beliefs about others. In Section Two, Wallace also attempts to fill in the Strawsonian gap. Wallace argues on behalf of a "disjunctive" account of moral responsibility, that the experience of the reactive attitudes is not, strictly speaking, necessary for holding responsible — only the judgment that the reactive attitudes *would* be appropriate *if* experienced. But, although Watson and Wallace are right to criticize Strawsonian minimalism on the grounds of explanatory inadequacy, and to emphasize the importance of history to the responsiveness of our attitudes, we will see that even they fail to take the problem of the causal origins of the RAs seriously enough.[1] I develop the explanatory turn further in Chapter Three, where I consider several non-Strawsonian hypotheses which, or so I will argue, *do* provide the best explanation for features of the RAs.

Section One: Watson Psychopathy and Child Abuse

Watson (Watson, Responsibility and the Limits 120) agrees that the RAs may not be as "innocent of theory" (his phrase) as Strawson's minimalism would suggest. For instance, Watson examines one crucial sort of exempting condition, psychopathology, which he thinks that Strawson's theory cannot adequately account for. I think that Watson is on to something

important in assuming that explanatory power is an integral part of any adequate account, and a weakness of Strawsonian minimalism. Although Watson is right to stress the role of *history*, by stressing the *history of the object* of the reactive attitudes, he neglects to consider the role played by the *history of the subject* experiencing them. In neglecting the history of the subject or person evincing the reactive attitudes, Watson also fails to find a satisfactory explanation for Strawsonian exemptions. In Chapter Three, I attempt to fill this explanatory gap. In the meantime, we will look at Watson's provocative discussion of the moral responsibility of the psychopath.

Drawing on Strawsonian philosophical anthropology, Watson distinguishes the two types of "pleas" that are said to inhibit or modify the negative RAs. His *type-1* pleas correspond to Strawson's category of standardly acknowledged *excusing* conditions. Persons may be excused by denying that the other temporarily failed to fulfill the basic demand for good will. For instance, if a potential subject of blame is ignorant, then the potential subject has not, in fact, failed to conform to the basic demand. Therefore, any resentment or blame will or should subside. Watson's *type-2* pleas represent Strawson's category of the standard *exempting* conditions like being psychotic or being a child. According to Watson's reading of Strawson, exemptions of the type-2 sort show that:

> ...the agent, temporarily or permanently, globally or locally, is appropriately exempted from the basic demand in the first place. Strawson's examples are being psychotic, being a child, being under great strain, being hypnotized, being a sociopath ("moral idiot"), and being "unfortunate in formative circumstances". His general characterization of type 2 is that they present the other either as acting uncharacteristically due to extraordinary circumstances, or as psychologically abnormal or morally undeveloped in such a way as to be incapacitated in some or all respects for "ordinary adult interpersonal relationships" (123).

In short, Type-2 pleas involve assessing whether the agent is an appropriate "object of that kind of demand for goodwill or regard which is reflected in ordinary reactive attitudes," including, as we can see, some consideration of an agent's history, or being "unfortunate in formative circumstances" (Strawson, Freedom and Resentment 9). Recall that, according to Strawson, if a person is exempt from moral demands, then we should take up the objective stance toward his moral failures and treat the person as an object of inquiry or therapy, as one to be "controlled, managed, manipulated and trained…" rather than as an appropriate object for resentment or indignation.

Watson thinks that Strawson's minimalist account is incomplete in that it does not fully explain pleas of the second type. The problem is this: RAs are said to depend on an *interpretation* of conduct. They are "natural human reactions to the good or ill will or indifference of others as displayed in their attitudes and actions" (Strawson, Freedom and Resentment 10). More specifically, negative RAs respond to the conduct and attitudes of one who flouts "the basic demand for reasonable regard" (122). But, Watson objects, *something more than this* must explain exemptions from the negative RAs since, "a child can be malicious, a psychotic can be hostile, a sociopath indifferent, a person under great strain can be rude, a woman or man "unfortunate in formative circumstances" can be cruel" (125). In other words, there are some persons who *do exhibit ill will* but who are nevertheless *exempted.* Consequently, Watson concludes that what it means to be exempt from the basic demand cannot amount to the mere presence of ill will (or the mere absence of good will).

Watson also argues that merely being "incapacitated from adult interpersonal relationships" cannot explain exemptions, as Strawson claims, since we obviously do not stop blaming some persons just because we wouldn't want a relationship with them. There

are some mass murderers, for instance, who do not have abusive histories, but who, nevertheless, lack communicability or respect for moral tenets. Strawson suggests that part of what it means to be suitable for adult relationships and moral demands is to share given ends or be a "member of the moral community" (Strawson, Freedom and Resentment 17). But, Watson insists, some bad apples lack membership in the "moral community" and the capacity for "adult interpersonal relationships," yet we thoroughly resent them. After all, the incapacity for such relationships, e.g., their lack of respect for our moral beliefs, their lack of civil responsiveness required for relationships may, in fact ground, our responses. In other words, the mass murderer's disregard for our moral standards, and his seeming incapacity for relationships, *intensifies* rather than *mitigates* our reactions of indignation. If being a suitable candidate for adult relationships and membership in the moral community adequately explains exemption from the basic demand, then, argues Watson, "some forms of evil will be exempting conditions" and the "paradox results that extreme evil disqualifies one for blame" (131). In sum, Watson rightly concludes that *something other* than the absence of good will or membership in the moral or inter-personal communities must explain exemptions from the basic demand.

One of Watson's main ideas is that the RAs seem to be sensitive to our beliefs about the *histories* of others. Strawson acknowledges this, maintaining that the RAs do respond to those having "unfortunate formative circumstances." Watson maintains that, given the *importance of personal history* to moral responsibility, we *ought* to be deeply worried about our lack of knowledge about how moral violators came to be as they are.

> Why does our ignorance not give us more pause? If, for whatever reason, reactive attitudes are sensitive to historical considerations, as Strawson acknowledges, and we are largely ignorant of these matters, then it would

> seem that most of our reactive attitudes are hasty, perhaps even benighted, as skeptics have long maintained. In this respect, our ordinary practices are not as un-problematic as Strawson supposes (145).

Watson points out that although the reactive attitudes respond to historical considerations, it remains unclear whether Strawson's account provides the conceptual resources to explain such historical exemptions. He uses the case of the psychopathic murderer Robert Alton Harris to dramatize the point.

Harris brutally murdered two teenage boys without provocation, and afterwards displayed heartless brutality by calmly eating his victims' unfinished hamburgers and making jokes about the murders. According to Watson, reading this sordid tale evokes strong indignation. But Watson notes that the very same person learning the facts of this crime, and who has just reviled Harris can almost sympathize with Harris upon learning further the sad facts about his terrible childhood. Although as an adult Harris displayed a willingness to inflict suffering in the most macabre fashion, we suspend our RAs in response to new information — the horrific abuse and neglect he suffered as a little boy.

For instance, we learn Robert was the premature baby of an impoverished alcoholic mother who delivered him after being kicked in the stomach by Robert's alcoholic, sexual-molesting father. We learn that little Robert once approached his mother "just to try to rub his little hands on her leg or her arm," only to be pushed away so violently that his nose became bloodied. Watson points out that once we encounter such facts about Harris's abusive childhood, the indignation we first experience wanes and a kind of sympathy develops. We think: "No wonder Harris is as he is!" (136). But Watson asks, "what is the relevance of this thought?" (137). Strawson must tell us more about *when* the RAs respond

to historical considerations. He must tell us more about why he thinks that "being unfortunate in formative circumstances" exempts some people from moral responsibility.

Watson does seem to acknowledge that our negative RAs toward Harris would subside, and sympathy, rather than antipathy, would gain a foothold, if we had known Harris as a child, or were able to keep his childhood vividly in mind (139). However, Watson stresses that our indignation does not completely wane as the incompatibilist determinist maintains. Rather, we experience emotional "ambivalence" toward Harris, which involves a shifting between "sympathy" for the boy he was, and "outrage" towards the man he has become (138).

Watson argues that Strawsonian minimalism has a hard time explaining our ambivalent responses to psychopaths like Harris. What is it about Harris' history that diminishes, or even temporarily, extinguishes our RAs? Is there any relationship between lack of membership in the moral and inter-personal communities, and the sensitivity of RAs to historical considerations like childhood abuse? Are we to regard psychopaths like Harris as *exempted* by virtue of his unfortunate history, or as *evil*, and flouting the moral and inter-personal relationships he possessed the capacity to have? Watson maintains that we typically do feel indignation and resentment toward Harris, which suggest his evilness. However, we also respond to historical pleas on his behalf, which suggests exemption. Watson suggests that Strawson's account might result in some forms of evil like Harris' — providing their own exemption from moral responsibility (132). Watson offers several additional 'candidate explanations' to explain moral exemptions, and complete Strawson's account, but to no avail. Watson speculates that a compatibilist-friendly explanation for the effect of our beliefs about personal histories on our RAs may come from further development of the concept of being a

fit candidate for "intelligible moral address" (148, f. 31). If, however, being the appropriate subject of "intelligible moral address" is a requirement for being subject to the basic demand for regard, and we define this elusive notion in terms of the idea of being willing to enter into a dialogue (134), then, yet again, some evil persons would inappropriately get off the hook. After all, there are evil persons who would not be subjects for "intelligible moral address" in that they reject the very premises of our moral community which make talk about morality possible. Watson admits defining the concept of "intelligible moral address" in a way that satisfactorily explains moral exemptions "requires further delineation" (148, f. 31).

Watson argues further that future theorists may provide an account of the compatibility of the reactive attitudes with the truth of determinism. But, since we do not typically exempt psychopaths like Harris because of his abused childhood or lack of membership in our moral and relational communities, we ought to be *skeptical* about Strawsonian minimalism: contrary to Strawson, philosophical anthropology does not reveal the genealogy of the RAs to be solely non-controversial beliefs. Minimal beliefs about the intentions and histories of others seem insufficient to explain some of our observations and experiences of the reactive attitudes; and it "remains unclear to what extent our ordinary practices involve dubious beliefs about ourselves and our histories (146).

But, Watson seems mistaken to think that Strawson cannot revise his account with an historical condition without conceding to the incompatibilist. Answering the question of just what historical conditions are required for moral responsibility is a daunting task. Nevertheless, there are compatibilists that are historicists.[2] As we shall see, historicists such as R. Jay Wallace focus their efforts on arguing against the incompatibilist generalization arguments. In suggesting that any appeal to a historical condition entails incompatibilism,

Watson merely assumes (rather than argues) that all historical conditions entail that determinism itself is an exempting condition.

As I suggested at the outset, implicit in Watson's conclusion is the *methodological assumption* that any adequate theory of moral responsibility must provide, among other things, a satisfactory *explanation* for our reactive attitude data. Watson argues, in effect, that we ought to be skeptical about Strawsonian minimalism because it *lacks explanatory power* and fails to adequately explain our responses to psychopaths, their historical pleas, and so on. Does R. Jay Wallace's historicist account of responsibility provide a more successful explanation for our reactive attitudes?

Section Two: Wallace on Exempting Normative Incompetence

Implicitly furthering the lack of explanatory force critique of Strawson is R. Jay Wallace. One of the important contributions that his *Responsibility and the Moral Sentiments* makes is to deepen our understanding of the Strawsonian conditions of appropriateness, especially the exemption condition. According to Wallace, we are prone to respond to such "unfortunate formative circumstances" as an abusive childhood because it undermines or reduces the victim's capacity for "reflective self-control." More specifically it undermines and reduces the development of the person's capacity to recognize, apply and act on moral reasons. According to Wallace, it is this, the capacity for reflective self-control, that is the "basic prerequisites for accountability" and that to which our RAs respond (215). The absence of, or diminished capacity for, reflective self-control renders it "extremely difficult" for the person to respond to the basic demand. Wallace argues that since "fairness" requires making the "difficulty of exercising one's general rational powers" an exempting

condition (224, f. 28), our negative reactions cease, or ought to cease, towards violators who have suffered abusive childhoods such as Harris'.

Wallace might explain Watsonian "ambivalence" toward Robert Harris in the following way. Upon learning about Robert's tortured childhood, we should suspend our indignation and moral disgust out of fairness, since he never had a real chance to develop the capacities, which ground moral capacities. After all, histories of childhood abuse diminish the capacity for reflective self-control, and makes acting morally extremely difficult (215). Contrary to the Watsonian critique, Wallace concludes that Strawson *can* use facts about an agent's history as exemptions since the incompatibilist generalization arguments fail. Determinism does *not* entail anything about history, that itself entails that an agent's capacity for reflective self-control is reduced or eliminated.

Wallace echoes Watson's worry about the explanatory inadequacies of Strawsonian minimalism. He argues the "sensitivity" of retributive reactions are *not* explicable *merely* in terms of beliefs about whether an agent's intentions and histories exclude him from moral and inter-personal communities. Wallace suggests moral responsibility is, what Fisher and Ravizza call, "genuinely historical", and not, as Strawson implies, merely "epistemically historical" (Fischer and Ravizza 173-194). For Wallace, moral responsibility is "genuinely historical", for it cannot, in principle, be explained *solely* in terms of facts about the *present* psychological states of agents. For Wallace, moral responsibility also requires abilities to grasp and apply moral reasons, including the "fundamental commitment to fairness", that are *themselves* historical notions (Wallace 102). Rather than agent history, such as experiencing an abusive childhood, as a mere *means* to some *current* feature about them, Wallace argues

such historical deprivation may show the capacity for rational self-control has been damaged, and acting according to moral reasons is, in varying degrees, difficult.

But, although Wallace makes the historical processes, identifying and acting upon moral reasons and being *held* morally responsible *necessary* for *being* morally responsible, historical facts are *not* sufficient. As we have seen, in his effort to avoid the source problem, Wallace makes *being* morally responsible, at least in part, a function of one's history. But, it is hard to say whether Fisher and Ravizza's notion of the "genuinely historical" fits Wallace's account — if not moral responsibility itself. After all, Wallace maintains being responsible involves the *fairness* of holding responsible. But, if *non*-historical conditions such as "fairness" is involved with being morally responsible, this means, two agents with the *same* degree of rational self-control, and, the *same history*, could in principle, *vary* in their responsibility. If normative issues such as fairness *are* involved with being morally responsible, it is hard to say whether moral responsibility is, as Fisher and Ravizza argue, an "essentially historical" notion (Fisher and Ravizza 207). More importantly, if being morally responsible *does not*, strictly speaking, require the past to be a certain way, then, how do we know the *source* of the agent's capacity for rational self-control (in virtue of which he is to be *fairly* held responsible) *is* his own, in some morally meaningful sense? Wallace's normative account of moral responsibility does not then avoid the problem of the "implantation" of desires or source problem.

Moreover, Wallace's version of historicism and retreat to normativity does *not* seem to help us with the problem posed by the seamy side of the reactive attitudes, that is, the *version* of the source problem emphasized here. He is however, highly sensitive to the sort of "emotional vicissitudes" which, I will argue, are a pervasive part of the reactive stance

(Wallace, Appendix 1, 237-250). For instance, Wallace seems to share my suspicions about the dubious genealogy of many (if not all) of our reactive attitudes:

> Far more interesting is the tendency of indignation to take on a harsh and punitive character, a tendency that can lead to forms of emotional *irrationality*. Indignation can be a way of channeling aggression onto others; indeed there may be an element of *aggression* present in *all* instances of indignation (and other reactive emotions). If this is right, then we can see how (as with guilt) an antecedent aggressive impulse might be the *original cause* of one's indignation toward a person. Indignation requires the belief that the targets of the emotion have violated some demand that we hold them to, but *the* reason we hold them to this demand in the first place might be a prior feeling of hatred for which we seek a socially acceptable outlet. A racist for instance, might come to feel indignant toward the blacks who live in his city for what he takes to be their slovenly habits, while the filth of the white workers' districts leaves him completely unmoved (Wallace 250, italics mine).

Here, Wallace acknowledges that "all" reactive attitudes may involve "aggression", and that we may hold persons to the moral demand "in the first place" because of some antecedent aggressive impulse. Although a white racist believes that slovenliness *is* a moral failing, he may only experience the feeling of indignation towards slovenly blacks — not slovenly whites. Wallace worries a lot about the apparently *inconsistent* nature of our emotional responses. As we have seen, in the attempt to avoid the problem of the dubious genealogies of the RA, Wallace adds a causal condition to the minimal conditions provided by Strawson. According to Wallace, we *can* distinguish rational from irrational or pathological forms of resentment and indignation since rational RAs are caused by the belief that a moral expectation one accepts has been violated — rather than something else, e.g., prior aggression or hatred (40-50).

Unlike Strawson, Wallace takes the "seamy side" of the reactive attitudes very seriously. However, although Wallace maintains that his robust causal condition renders the

RAs *immune* from dubious psychological and otherwise irrational causes, e.g., "self-deception", "guilt transference" and "sadism" (Strawson 26), his version of historicism does *not* help Wallace *dialectically* — even *he* does not take the reactive data serious enough. After all, Wallace does not satisfy *the pessimist* who longs for that "vital thing" missing from the consequentialist conception of responsibility. Wallace acknowledges that the RAs are culturally contingent:

> The reactive attitudes are not co-extensive with the emotions one feels towards people with whom one has interpersonal relationships... nothing in the very idea of this quasi-evaluative stance rules out the possibility of cultures whose members are not subject to the distinctly reactive emotions (31).

Wallace admits that some cultures (cultures that are *not* necessarily "psychologically and morally primitive" (31)) *include* moralities and inter-personal relationships but *lack* the reactive attitudes. But, Wallace argues that morality and inter-personal relationships do *not* depend on the actual experience of the reactive attitudes — even in this culture. Wallace worries so much about the "seamy side" of the reactive attitudes and notices the apparent inconsistencies of our reactive responses acknowledging that "...it is not the case that we actually feel the relevant emotion..." in all the cases we believe others have violated the basic demand. For instance, although we may believe a "charming acquaintance" *is* morally responsible for violating an expectation we accept, e.g., to refrain from lying, we may, nevertheless, fail to adopt the reactive stance towards her. But, Wallace maintains, in spite of the *absence* of such reactive feelings, we *can* hold our charming acquaintance responsible. For Wallace, although our culture possesses the concepts relevant to the RAs, moral

responsibility *requires no more* than the objective stance and the belief that the RAs *would* be appropriate, if they were experienced (31).

As we can see, Wallace takes the problem of the causal origins of the RAs seriously. He notices the *contingent* connection between the moral judgments we make and the reactive attitudes we actually experience, and makes only the former necessary for moral responsibility. He does *not*, however, review the data closely enough and offer an adequate explanation for *why* these variations occur. Wallace maintains that possessing a certain "resentful temperament" explains why some persons are prone to adopt the RAs while others do not (249). But antecedent hatred and the nature of temperament fail to account for such variations as why we react to some liars, but *not* our charming colleague. Although Wallace takes the problem of the causal origins of the RAs seriously, he fails to take the data seriously enough. He fails to articulate only *the* most important aspects of reactivity which jeopardizes the whole Strawsonian reconciliation strategy, namely, emotional *akrasia*.[3]

Wallace defines what it is to accept a moral demand *vis-a-vis* the "consistent" motivation to act and criticize others accordingly (41). The person who fails to react to a charming colleague believes that an expectation he accepts has been violated, and if asked, he would *judge* the liar negatively. But, like the racist, this person does *not* adopt the reactive stance consistently. Describing the manifestations of such a phenomena is tantamount to the philosophical difficulty of describing (let alone explaining), *akrasia* in general.[4] Suffice it to say however that the colleague *judges* that the attractive colleague has committed a wrong *and* that he himself *ought to feel* indignant towards her, but, he fails to do so.

Wallace fails to take the variations between our moral judgments and reactive responses seriously enough, since phenomena such as *akrasia* dull the luster of what Strawson hopes is the "glittering prize" of the retributive attitudes. Recall from Chapter One that the RAs are supposed to fill the void of "vitality" that the pessimist thinks is missing in consequentialist accounts of moral blame and punishment. To the extent he recognizes vindication of the reactive attitudes requires additional causal conditions, Wallace's historicist account seems to be on the right track. To be sure, Strawson wants to avoid specifying additional causal-historical conditions for he then faces the daunting task of denying the incompatibilist generalization arguments. But even if causal historical conditions alleviates some worries about *akrasia* or the inconsistencies of our emotional responses, any normative account of what ought to be the case does not help Strawson with the problem his reconciliationist strategy currently faces. After all, *the pessimist* must be persuaded that what *is appropriate* and "vital" to moral responsibility *is* grounded in what *is* the case, namely, the anthropological "fact" that to a "great extent" reactive attitudes *are* caused by minimal beliefs and attitudes about the intentions of others (Strawson 5). But, in allowing that the *experience* of the reactive attitudes is *never*, strictly speaking, necessary for holding responsible — only the judgment that the reactive attitudes *would* be appropriate if experienced — Wallace's "disjunctive" account *de-emphasizes* the reactive attitudes as they *actually are*, in the realm of human inter-personal experience. His treatment of reactive *akrasia* dulls the luster of the RAs because moral *judgments* become *the* central features in holding responsible — *not* the experience reactive attitudes themselves. Since the latter are dispensable and the former are not, Strawson's RAs become *dull and unimportant* — not

"vital"; the pessimist will remain unconvinced that what *is* appropriate to moral responsibility *is* to be found in the natural realm of anthropological "facts".

But, perhaps manifestations of the self or reacting subject such as emotional *akrasia* do *not* taint the RAs in any significant way. Strawson might respond that while episodes of emotional *akrasia* undoubtedly occur, they are *not* pervasive. A judgment based morality like Wallace's need not de-emphasize the experience of the reactive attitudes since "it is an exaggerated horror itself suspect, which would make us unable to acknowledge the facts because of the seamy side of the facts" (Strawson 24). But, as we shall see, emotional *akrasia* is *not* the only "seamy" side of the reactive attitudes. We turn now to further worries that the other directed retributive attitudes are all *too frequently* grounded in unacknowledged guilt and partiality.

Notes To Chapter Two

1. On the one hand, Strawson claims he is not concerned with the "general causes" of the reactive attitudes (Strawson 6). On the other, we shall see in Chapter Four that Strawson's account does involve controversial assumptions about general causes.

2. See Fisher and Ravizza and Michael McKenna for compatibilists that are historicists.

3. According to one definition, *akrasia* is the state in which humans lose power over themselves without obvious exterior agencies forcing them to do so; knowing the good, seemingly able to do the good, yet, doing something other than the good. *Akrasia* is said to pose a problem for ethics more generally; "Despite the alluring grandeur of its ideal of self-mastery, ethics has a skeleton in its closet, a dark secret in an otherwise perfectly ordered house. Socrates saw the problem at the very inception of ethics, but denied that it really existed. Aristotle attempted to explain it away, but it remains a genuine difficulty in his otherwise coherent system of ethics. The great thinkers of the Enlightenment were sure that reason could overcome it, but the problem did not go away. Rather, it exploded upon the West with such great force during the nineteenth century that we are still reeling from its implications. The Greeks called the problem "*akrasia*"" (Riker 41). As we shall see in Chapter Three, whether variations between mental judgments and reactive responses are best explained by *akrasia* or, something else, it is skepticism about the variations themselves that pose a problem for Strawson dialectically.

4. Marx for instance, believed *akrasia* is not an anomalous state for human beings, but rather, forces beyond their power and awareness constantly control humans. Regardless of virtues, reasoning, social norms, or desires (conscious or unconscious), for Marx, *akrasia* is best explained by the determining force of material and economic conditions.

CHAPTER THREE

The Seamy Side of Retributive Reactivity

Watson and Wallace seem right to criticize Strawsonian minimalism on the grounds of explanatory inadequacy, and to emphasize the importance of *history* to the responsiveness of reactive attitudes, but as discussed in Chapter Two they fail to take the problem of the causal origins of the RAs seriously enough. In Section One of this Chapter, I argue that there are important *variations* in the nature and existence of reactive affect that Strawsonian conceptual resources and cognitive conditions *fail* to explain, even when augmented with the materials Watson and Wallace provide. What are these unexplained variations? Strawson's minimalist belief conditions do not seem to account for important variations in some of the constitutive aspects of reactive affect, namely, D.I.R.E.: the *Duration*, *Intensity*, *Responsiveness* and *Existence* of reactive affect. What might provide a better explanation for the nature and existence of the retributive reactive attitudes then Strawsonian minimalism? Conceptual analysis and philosophical anthropology alone have hitherto failed to identify, much less explain, these important features. Instead, I develop the explanatory theme further and consider several non-Strawsonian hypotheses which, or so I will argue, *do* provide the best explanation for features of the RAs.

I argue *over and above* any beliefs about others, on many occasions important aspects of our RAs are *best explained* in terms of the "seamy side" of the retributive RAs involving certain unacknowledged beliefs and attitudes the reactor has towards him or herself. I offer empirical data supporting the notions that many, if not all, RAs are caused by what amount to dubious epistemic and moral beliefs and attitudes about *our own* history and interests.

I continue the argument in Section Two, maintaining that *one way* the self-influences the duration, intensity, responsiveness — even the existence — of the retributive attitudes he adopts towards others is the *projection of guilt*. I argue this neurotic (if not pathological) version of the "seamy side" is problematic *because* the reactor believes his or her retributive attitudes are *solely about others*. But if the retributive attitudes also presuppose the reactor's reflexive moral beliefs, then they are *not solely about others*. Therefore, the retributive attitudes that involve the projection of guilt *are* epistemically flawed, involving, as they do, false beliefs about their genesis.

I *extend* this argument, in Sections Three and Four, that over and above the glaring neurotic or pathological versions, there *are more modest versions* of the retributive attitudes that are "*seamy*" *nonetheless*. Aspects of the self such as beliefs and attitudes about one's proximity to the perpetrator or person harmed, personal history, ethnicity, religion, shared victim hood, and so on, *also* influence the retributive attitudes we adopt. I will argue that positing these modest reflexive beliefs and attitudes within the reactor, *frequently* provides a *better explanation* than *Strawsonian minimalism alone* for our retributive attitudes. These modest (and pervasive) retributive attitudes *are* part of the "seamy side" because, like their more exotic counterparts, they involve *lapses of self-understanding and impartiality*. The pervasiveness of the "seamy" retributive attitudes, i.e., of glaring *and* modest varieties, raises deep worries about Strawson's reconciliation project; the RAs may involve too many dubious beliefs to provide sufficient temptation for the incompatibilist to surrender metaphysics.

Section 1: Explanatory Inadequacies of Strawsonian Minimalism (D.I.R.E.)

What is the D.I.R.E. nature of reactive affect and how is it best to be explained? I will start by exploring the last of these features, namely the very *existence* of reactive affect. Suppose that there are graphic stories on the nightly news about murderers like Ted Bundy, Pol Pot and Joseph Stalin, which cause me to believe that all of them are equally responsible for violating the basic demand. Do we always, do we even typically, actually feel indignation towards all the offenders on the nightly news? Suppose that as I sit before the television set I feel indignation only towards one or perhaps two of these moral offenders, or that I fail to adopt the reactive stance towards any of them. As we have seen, Strawson suggests that negative retributive RAs like indignation are grounded in the general expectation that all persons must meet the basic demand for good will and regard. Once we come to believe that a particular offender in fact failed to adhere to the basic demand, and, he is neither excused nor exempted, then we typically come to adopt the retributive stance towards him. However, the variance in our affective responses to the three mass murderers on the nightly news reveals that merely embracing these two beliefs is not always enough to determine the existence of indignation towards them. We may feel indignation towards only one or some of those we believe to be offenders, or we may feel no indignation at all.

P.F. Strawson attempts to explain this phenomenon, i.e., failing to adopt the reactive stance towards those we believe are responsible for violating the basic demand (Strawson 6-7). He would argue that watching the nightly news could be an exhausting emotional roller coaster after a hard day of relating to others. Although we may believe that the criminals have violated the basic demand, and are neither excused nor exempted, e.g. they come from backgrounds compatible with membership in the moral community, we might not feel

indignation towards them all. What is Strawson's answer to this quandary? Strawson argues that we sometimes refrain from adopting the reactive stance out of a kind of fatigue. Strawson rightly suggests that it is an empirical *cum* psychological fact that human beings simply cannot react to all moral infractions, since doing so would be an unbearable tax on the limits of our emotional resources. The realm of the objective stance provides a vacation, of sorts, from the stress of inter-personal relationships, since it is by its very nature free of reactive affect. Adopting the reactive stance towards every moral violator would leave us emotionally exhausted and almost incapable of getting on with the other activities of life. Thus, Strawson might conclude, we may not react or react to only some of those we believe have violated the basic demand because the (impartial) over-all economy of our inner resources or random emotional fatigue engenders the retreat to the objective stance. The existential reactive data *are* accommodated by his conceptual toolbox.

This Strawsonian explanation for the variance in our interpersonal reactive affect is fine as far as it goes, but it does not go nearly far enough. For there are clear cases where emotional fatigue and the need to relieve the strains of involvement are not the issue. I maintain that the problem of specifying affect variation *includes* the need to explain why we *do* react to *J*, but *do not* react to *P or S*, given our beliefs that *all* are responsible for violating the basic demand for regard. Why, for instance, do *I* react with indignation towards Ted Bundy but not towards Pol Pot, or Joseph Stalin? Why, on the other hand, might *you* react indignantly towards Pol Pot and Joseph Stalin, but not Ted Bundy? And why, might *Z* react indignantly to the news anchorman's glibness, but none of the mass murderers he is being glib about? The retreat from exhaustive inter-personal relationships may explain why we feel indignation once, rather than three times, or why we can work up a lather towards some,

but not all, of those whom we believe to be moral violators. But, needing a break or retreating from the stresses and strains of inter-personal relationships does not satisfactorily explain why I experience affect towards Ted Bundy, but do not experience affect towards Pol Pot and Joseph Stalin. Nor why, on the other hand, you experience affect towards Pol Pot and Joseph Stalin, but not also Ted Bundy.

In addition to those unexplained inter-violator discriminations we make among the heinous, we also discriminate among moral violations committed by those we love in a way that is not always easily explainable in Strawsonian terms. I may discover, for instance, that a current lover fails to recycle and neglects the environment. But although I believe he *is* morally responsible for harming the interests of human beings and violating the basic demand, I fail to adopt the indignant stance towards him. Upon discovering, however, that a former lover also fails to recycle I *do* experience indignation towards him. Why might this be?

In cases like my adopting the indignant stance towards my former but not my current lover, this variance in our retributive RAs does seem to involve our own psychological needs or *egos*. It seems to have some effect on whether I adopt the reactive stance for instance, that *I* am currently in a romantic relationship with A but no longer with B, even though they both commit the same offense. Since I *do* consider my current lover as a member of the moral community and am *not* under strain, taking the objective stance towards harms perpetrated by others does seem to involve other than Strawsonian excuses, exemptions and exhaustion.

Moreover, there seem to be certain kinds of violations that engender RAs no matter who commits them, but other kinds which leave us unmoved, even though we do concede the equal moral seriousness of the violations. For example, I may always feel indignant towards

people who fail to make generous donations to the local food bank, but not towards people who fail to recycle, even though I firmly believe that degrading the environment poses a greater harm to humanity than neglecting the local food bank.

What explains the nature and existence of such "inter-violator" and "inter-violation" discriminations? Am I merely too *tired* to react towards my new lover for both his moral crimes? I think not. I am *not* tired, but full of energy and respect. I am eager to engage in discussions and emotional intimacies. Something over-and-above the Strawsonian desire for reprieve *is* needed to explain the variance in our retributive reactions to different people and to different kinds of actions.

Strawsonian minimalism seems to provide inadequate explanation for the *existential* variety of what seem to be many common place RAs. Further mysteries appear when we notice that the *duration* and *intensity* of RAs also vary in ways that seem unrelated to the desire to relieve the strains of involvement. One might resent both *S* and *T* for hurting one in the same sort of way, but one's resentment towards *S* might *linger far longer* or be felt with much *greater force*. For instance, our drycleaner and a colleague at work may both fail to honor our request to make a particular, but relatively unimportant deadline. Our resentment towards our colleague may endure, however, with an intensity that surpasses both the minor nature of the breach and the fact that our resentment towards the drycleaner for a similar offense was brief and mild. Why such variance? Nor does Strawson's notion of taking a reprieve from the strains of involvement does not explain why the resentment I feel towards *S* for insulting me, is more *enduring* and *intense* than the resentment I feel towards the C.E.O.'s of transnational corporations for spewing carcinogens into the air. Why is it that by the time I finish the morning paper, my mild resentment towards the C.E.O.'s subsides,

while, on the other hand, my resentment towards *S* for calling me a name endures, in moderate proportions, for days? If I believe that the polluter has committed a greater violation of my interests than the name-caller, then, what explains the greater *duration* and *intensity* of my resentment towards the latter? Furthermore, why does the *responsiveness* of our RAs to facts about a violator's history so often vary with the violator? That fact that S endured an abusive childhood may quickly dim my resentment, but not the fact that T suffered similar abuse. Why? Yet again, the Strawsonian explanation in terms a desire to relieve the strains of inter-personal commitment is inadequate to explain the variance.

Section Two: The Projection Hypothesis:

> No one is more ferocious in demanding that the murderer or the rapist 'pay' for his crime than the man who has felt strong impulses in the same direction. No one is bitterer in condemning the 'loose' woman than the 'good' women who have in occasion guiltily enjoyed some purple dreams themselves. It is never he who is without sin that cast the first stone. Along with the stone, we cast our own sins onto the criminal. In this way we relieve our own sense of guilt without actually having to suffer the punishment — a convenient and even pleasant device for it not only relieves us of sin, but makes us feel actually virtuous (Weihofen 138).

Is, as Weihofen suggests, the D.I.R.E. of our retributive reactions, to "murderers", "rapists", "loose women" and so on, explainable in terms of the contingencies of the evincing self? We have reviewed some of the difficulties that Watson and Wallace think that Strawson's minimalist belief thesis have in accommodating important aspects of our reactive experience. I have maintained that implicit in Watson and Wallace is what we might think of as an explanatory turn in the responsibility debate; rather than relying on conceptual analysis and philosophical intuitions, any adequate theory of reactive responsibility must provide the

best explanation of certain features of our RAs. What then might explain our reactions to psychopaths, childhood abuse, the nightly news and former lovers? In this section of Chapter Three, I develop the explanatory theme further and consider several non-Strawsonian hypotheses, which provide, I will argue, the best explanation for features of the RAs. Conceptual analysis and philosophical anthropology alone have hitherto failed to identify, much less explain, these important features.

To bring these D.I.R.E. themes together, consider Sarah, a college coed whose financially-strapped parents consistently remind her of the severe financial hardship that parenting imposes. Sarah's parents frequently berate her for her wastefulness and teach her that lessening their economic burden is among the highest of moral obligations. But, in addition to scolding, Sarah's parents also provide her with support and encouragement. Although she imposes financial hardships, Sarah's parents tell her that she is charming and funny and that she is admirably honest and particularly trustworthy, unlike many teenagers. So, like many of us, Sarah believes she has fallen short in some ways, e.g., for failing to address her parent's poor financial situation, but has achieved moral integrity in others, e.g., through behaving in a trustworthy and honest fashion.

Upon entering college, Sarah learned, among other more pleasant things, that a coed was brutally stalked and assaulted, that a classmate submitted essays he purchased elsewhere, and that she failed to win a scholarship competition she had so diligently pursued. What is the D.I.R.E. nature of Sarah's intra-violator reactive affect? Suppose that Sarah does not feel any particular indignation towards the cheater, only brief and mild indignation towards the coed attacker, and intense and enduring indignation towards the winner of the scholarship, a

young woman named Hillary. What factors might explain the D.I.R.E. variance in Sarah's RAs?

First, we notice what specific propositional form Sarah's reactions take. The RAs are intentional states with propositional objects. For instance, Sarah's indignation towards Hillary involves the proposition "that Hillary celebrated winning the scholarship by throwing a decadent party." Although winning a scholarship is not, she supposed, a moral crime, Sarah believed that Hillary failed to fulfill the basic demand by celebrating her good fortune in such a 'decadent' way. Sarah claims that she feels such indignation towards Hillary because "Hillary did not share any of the proceeds from her scholarship with the poor in the Developing World, but instead threw a wild and wasteful party." Although Sarah *also* believes that the cheater and the attacker have violated the basic demand, *why* does she feel no indignation towards the cheater and only mild indignation towards the attacker? This seems odd because Sarah would grant that sexual assault is a much worse violation of the basic demand than neglecting the poor. The enduring temporal character and stomach churning severity of Sara's sense of indignation towards Hillary seems to involve *other than* her *impartial concern* about Hillary's neglect of the world's poor. What, then explains these types of reactive affect allocations?

On the face of it, the beliefs and attitude conditions in Strawsonian minimalism — including his notion of emotional weariness — fail to account for these features of reactive experience. Can we find an explanation for the D.I.R.E. nature of Sara's reactive experience? Over and above any beliefs and attitudes Sarah might hold about *others*, it seems to me that a fuller explanation of reactive attitude D.I.R.E. posits the existence of certain beliefs and attitudes that Sarah holds about *herself.* How does the proneness to adopt

reactive attitudes, reactive attitudes that is, that are purportedly about *others*, involve the reflexive attitudes of their *subject's*?

One way the RAs presuppose their subject's reflexive attitudes focuses on the notion of *guilt projection.* Recall that Sarah feels sustained intense indignation towards Hillary for Hillary's failure to use surplus money to feed the poor (*p*). Sarah does not, however, feel sustained intense indignation towards the plagiarist for his dishonesty (*d*) — even though she consciously avows moral distaste for both negligence and dishonesty. And although she experiences some indignation towards the attacker for brutal assault (*b*), this token of Sarah's indignation is mild in intensity and short in duration. If Sarah believes that the dishonest cheater has violated the basic demand and is not excused or exempted from forging essays, why does she fail to adopt the indignant stance towards him? As discussed, Strawsonian minimalism fails to provide adequate resources to explain why Sarah *does* experience indignation towards Hillary, but *does not* towards the cheater. Also recall how intensely Sarah's parents complained about her wastefulness and neglect of their economic well-being. It is psychologically realistic to suppose that such moral badgering has caused Sarah intense feelings of guilt for these faults (whether real or imagined). If Sarah feels strong guilt for neglecting the financial hardships of others (minimal guilt for committing physical violence, and no guilt at all for cheating), then we *can* explain the discriminations of Sarah's inter-violator D.I.R.E., that is, why she feels such intense and enduring indignation about Hillary's neglect of the developing world's poor, only mild and fleeting indignation towards the attacker, and none at all towards the cheater. Note that Sarah does *not* feel intense guilt about neglecting the poverty in the developing world *per se*. She *does* feel intense guilt, however, about neglecting the poverty of her parents. In this case, the nature of Sarah's

antecedent belief that she is guilty is neither straightforward nor obvious, in that Sarah does not believe that she too has committed *p* itself, that is, has neglected the *poor in the developing world.* However, Sarah's projected guilt is about some thematically related *q*, namely, neglecting her own *poor and overworked parents*. The target of her most intense guilt, i.e., her own 'neglect of her poor parents', is *thematically much more similar* to the target of her most intense indignation, i.e., 'Hillary's neglect of the poor (in the developing world)', than to the dishonesty of the cheater or the brutality of the stalker. In other words, the nature of the *self's* reflexive beliefs, influences the nature of the RAs adopted towards *others*. The projection hypothesis suggests that rather than *merely* responding to the intentions and histories of others, the RAs also respond, at least in part, to the dysfunctional mechanism of pain-alleviation of denied guilt. The projection of guilt provides a *better* explanation than Strawsonian minimalism for why Sarah is more prone to intense and enduring indignation towards Hillary than towards the other offenders, even though she would be prepared to grant that self-indulgence is a less serious violation of the basic demand than either dishonesty or brutal assault.

After all, we tend to avoid pain. Guilt is psychologically painful. There are many ways to rid oneself of the pain of guilt. Sometimes we may try to forgive ourselves or make amends and restitution to the person one feels guilty about. However, sometimes these sources of relief are unavailable, for example, when one's guilt is unreasonable or when one's self esteem is low. It seems that the lower one's self-esteem, justifiably, the lower one's capacity to bear still further self displeasure of facing one's moral shortcomings. Sarah is actually quite sensitive to her parent's material needs. When she was living at home she always took summer jobs to help ease their financial burdens, learned how to handle money

prudently, and so on. But even though her parent's complaints are quite exaggerated, Sarah nonetheless thinks of herself as a spendthrift. Her guilt in this area is heavy. She tries to relieve it by cutting back even more in her spending, by going to a local college rather than the out of province school she would prefer. But nothing is quite enough for her parents. They continue to badger her about her material neglect of their well-being.

How might Sarah find some relief? A common defense against psychic pain involves the projection of the real or imagined flaw onto another person.[1] And so it is with Sarah. Although she dislikes having lost the scholarship to Hillary, this turn of events can become psychologically expedient for her, because it gives her an opportunity to focus on or scapegoat someone else's "extravagance" and bask in the considerable pleasures of self-righteousness or increased self esteem. For instance, instead of saying to myself, "*I* hate myself," I can say, "*He* hates me." Instead of saying, "*My* conscience is bothering me," I can say, "*He* is violating my interests." In essence, Sarah projects her guilt onto Hillary, thereby avoiding the pain of self-confrontation and moreover, the anxiety or fear of punishment that accompanies her guilt. Psychological projection, or some such functionally equivalent mechanism, relieves anxiety and fear by *switching the moral* subject. "*I* am morally responsible for neglecting my poor parents" gets switched into "*Hillary* is morally responsible for neglecting the poor in the developing world." More generally, projection mechanisms switch "*I* am guilty for *p*" for, "*She* is guilty for *p* (or some thematically related *q*)", and, "I deserve punishment for *p*" becomes "*He* deserves punishment for *p* (or some thematically related) *q*", and so on.

Why might Sarah's projected guilt be so very intense and enduring? Here, we must speculate a bit. Children cannot recognize their exemption from the basic demand and the

irrationality, or any pathological genealogy, of the negative RAs, which adults direct towards them. After all, to do so involves embracing the frightening belief that there *is* something seriously wrong with their parents, teachers, clergy and so on. But, children resist such a belief since it is tantamount to facing death itself, if there is something wrong with their parents, then what will happen to them? To accept that the resentment directed towards them may be rooted in their parent's needs, e.g., for bolstered self-esteem and the avoidance of guilt, is to accept the egoism and fallibility of those upon whom children absolutely depend. It is therefore a lot easier for children to believe that their parents are right about them, and they *are* morally responsible and guilty for violating the basic demand.[2] For example, in Sarah's case, the parental resentment directed towards her for her purportedly "spendthrift" ways and "negligence" does, in fact, involve Nietzschean mechanisms and antecedent guilt. As it turns out, in order to indulge small luxuries like dinning out and buying theatre tickets, Sarah's parents choose a sub-standard retirement home for *their* elderly parents. In resenting Sarah, her parents unconsciously sought the avoidance of painful guilt, namely that *they* are spendthrifts and are guilty for neglecting *their* poor parents. Like all children, however, Sarah finds herself caught between the horns of a painful psychological dilemma; she needs to believe that her parents are right and that there *is something wrong with her*, because doubting them involves believing something worse, namely, believing that there *is something wrong with them*. But if there is something wrong with her parents, in fact so severely wrong that they would scapegoat their little girl for relief, Sarah must confront deep anxieties about parental rejection and her own existential vulnerability. Since young Sarah's life depends on parental care and protection, rather than believing that there is something wrong with them, for instance that they are irrational, Sarah opts for the first horn of the dilemma and comes to

believe that the judgments contained in the RAs are right; she *is* guilty. However, since a child's very life depends upon maintaining the approval of her caretakers, she rejects the other horn of the dilemma, namely facing their (projected) guilt and disapproval can be tantamount to facing death. Unconsciously, through Nietzschean mechanisms, Sarah finds a way around the parental reactive attitude dilemma. On the one hand, Sarah "accepts" her guilt and preserves her faith in her parent's competence and love; on the other, she avoids the feelings of rejection and abandonment that this involves, by repressing the guilt and avoiding its pain by projecting it onto others.

Notice, too, that the projection hypothesis can also explain why Sarah feels some indignation towards the coed attacker for committing brutal physical assault, even though Sarah has never committed a brutal physical assault herself. Her indignation may, nonetheless, be a projection if she does unconsciously believe she is guilty for committing some transgression, which is *sufficiently related thematically* to physical assault. Neglect of her parents' well being might be perfectly fitting here, or perhaps some painful guilt about her rough and tumble treatment of a younger sibling. The reflexive beliefs and attitudes of the subject, in this case Sarah's guilt, seem to provide greater explanatory power than Strawsonian minimalism.

Moreover, still more D.I.R.E. variance is explainable in terms of the reflexive beliefs of its subject. If Sarah unconsciously believes that she deserves compassion or, at worst, brief and mild sanctions for her transgressions, then we would predict that she will be compassionate, or feel only brief and mild indignation, towards the indiscretions of others. On the other hand, if Sarah's unacknowledged beliefs about her own guilt include a harshly retributive notion of punishment, then, as we would expect — the target of her projection,

Hillary has reason to fear. For if Sarah's reflexive beliefs include the notion that neglecting her parents involves punishment, e.g. that she deserves torture or hell then, this may in turn dispose her to believe that Hillary deserves the same for *her* neglect of the poor. In other words, Sarah's reflexive moral belief, "*I* deserve the torture of hell for *my* negligence," becomes "*Hillary* deserves the torture of hell for *her* negligence." If Sarah lacks the conceptual and emotional resources to achieve deeper self-knowledge and an integrated personality, and rid herself of her own guilt about neglecting her parents, she must accept and confront the pain of what she thinks she deserves, namely the torture of burning in hell. In any case, the greater the *duration* and the *intensity* of the punishment, which "Draconian" Sarah believes she deserves, the more urgent the psychological "proneness" to adopt RAs towards others of the *enduring* and *intense* sort.

In sum, why, then, does Sarah adopt intense enduring indignation only towards Hillary for purportedly neglecting the poor, whereas her indignant stance is fleeting or nonexistent towards the other moral violators? The explanation for this lies in noting that Sarah enjoys a healthy self-esteem... at least with respect to her virtues of honesty. Sarah therefore, experiences not an ounce of unconscious guilt about her own honesty. She is not esteemed, however, by the belief that she has neglected her poor parents. Consequently, with no unconscious anxiety, pain or fear to relieve about being dishonest, but possessing unconscious anxiety, pain or fear to relieve about neglecting the poor, Sarah does not need to project indignation towards the cheater. She does, however, need to project indignation towards the neglect of the poor. Focusing on Hillary's squandering ways enables Sarah to distract herself from the pain of her own (imagined) extravagance since Sarah's attention is diverted and her guilt gets transformed into enduring and intense indignation toward

Hillary's (alleged) extravagance. However, Sarah feels no particular pressing guilt about issues of dishonesty or violence. Pain relief is available then *vis a vis* adopting the reactive stance towards Hillary. Sarah therefore, is less prone to feel indignation toward the cheater or the stalker than Hillary. Of course, she may well judge them as worse moral offenders than Hillary, but she does so (we have supposed) without much reactive affect.

Where then, does this analysis of the dubious genealogy of D.I.R.E. leave us? I take this inter-violator discrimination scenario to be perfectly realistic. I have argued that over and above any beliefs and attitudes a reactor holds about *others*, a fuller explanation of reactive attitude D.I.R.E. posits the existence of certain beliefs and attitudes that a reactor holds about him or *herself.* We have seen that *one way* that the RAs seem to presuppose their subject's reflexive attitudes involves the notion of *guilt projection.* We frequently observe those whose reactions to moral violators seem to involve the subject's guilt, or some other unmet psychological need or attitude. It seems that just as some passionate desire for the charming colleague seemed to inhibit or crowd out the emotional aspect of the moral judgment about the colleague's lying, so also Sarah's motive to rid herself of guilt seemed to inhibit or crowd out the emotional aspects of her moral judgment of the coed stalker. While moral judgments remain constant, the character and existence of the reactive emotions are disturbingly contingent.

Now, the emerging problem is this: If I am right and beliefs and attitudes that the *subject* of the RAs adopts towards his or her own actions and history play a powerful role in common place other directed RAs, then, it is pretty clear how this *non-Strawsonian* explanation tends to undermine its rationality. After all, Sarah *believes* that the duration, intensity, responsiveness and the very existence of her indignation towards Hillary is *solely* a

function of her beliefs *about Hillary*, i.e. Hillary's violation of the basic demand. But this, Sarah's "meta-belief" about the origin of her indignation is deeply mistaken. Sarah is in the grips of unconscious guilt. She succeeds in relieving the pain by projecting it onto Hillary in the form of intense, enduring indignant affect. If Sarah's indignation is *not* solely about the object of her indignation, i.e. Hillary, but *also* its subject, i.e. Sarah herself, Sarah's indignation is irrational, it is commonly accepted that if an emotion is grounded upon false beliefs, then it is criticizable from the standpoint of theoretical rationality.[3]

But if the nature of the self's reflexive beliefs influence, the "proneness" to adopt the other directed RAs that presuppose false beliefs, *is* this connection "pervasive"? Do such responses as *akrasia* and the projection of guilt play a powerful explanatory role of the ebb and flow of inter-personal relationships and the practice of moral responsibility, or, are our worries merely "exaggerated", as Strawson suggests? (25) While there may be *some* connection between the self's reflexive beliefs and attitudes and the retributive attitudes adopted towards others, is this connection *sufficiently commonplace* to be worrisome? If there is a pervasive causal relationship between reflexive beliefs and attitudes, on the one hand, and the retributive RAs we adopt, on the other, then the rationality of the retributive RAs are more compromised than Strawson acknowledges. Therefore, we must speculate a little about just *how* pervasive this "seamy side" of the retributive reactive stance actually is.

Section Three: Projection as One Breach of Impartiality

In the first two sections of this Chapter, I argued that Strawsonian minimalism, the thesis that the RAs *typically* presuppose merely non-controversial beliefs and expectations about the intentions of agents, provides insufficient resources to explain many important

variations in the nature and existence of the other-directed retributive RAs. Now, any claims about what is *typical* or *most frequent* in the reactive sphere involve *empirical* issues, a difficult subject matter for philosophy. But Strawson reminds us that although philosophy is a "theoretical study", we must also take account of the empirical facts, "in all there bearings" (25). Here then, I try to strengthen my general claim that the retributive RAs are *more compromised* by false beliefs than suits Strawson's *dialectical purposes*, by extending my argument to include psychological mechanisms other than *akrasia*, and the reactor's projected reflexive moral beliefs and attitudes. I shall argue, that the reactive stance is rife with failures of self-knowledge or self-understanding mainly attributable to *lapses of impartiality*. Moreover, I shall also argue that this problem threatens Strawson's reconciliationist strategy of offering the incompatibilist a tempting natural "vital thing" in return for his "panicky metaphysics" of the self. As noted in the Introduction: "tainted goods" tend to make the buyer beware.

As discussed, Strawson *does* acknowledge that the reactive stance has what he calls its "seamy side" (25). The key passage is worth quoting again:

> ...psychological studies have made us rightly mistrustful of many particular manifestations of the attitudes I have spoken of. They are a prime realm of self-deception, of the ambiguous and the shady, of guilt transference, unconscious sadism and the rest. But it is an exaggerated horror itself suspect, which would make us unable to acknowledge the facts because of the seamy side of the facts (Strawson 25).

As we have seen in Section Two of this Chapter, one aspect of this "seamy side" involves psychological mechanisms like the unconscious projection of guilt. Strawson himself refers to such mechanisms as "guilt-transference" and "unconscious sadism" above and admits that discovering such tainted RAs has increased the "difficulty of (their)

acceptance". Strawson is right, however, that such neurotic, indeed pathological, psychological features *do not* ground all of our RAs. He is also right that these features *do not* so pervade our RAs that they lose all their import for the practice of responsibility (24-25). However, the point here is this: Strawson *is* too complacent about the pervasiveness of more homely and therefore more frequent lapses from the impartial point of view. Here, in Section Three, I will argue that in addition to the ignorance, self-deception and irrationality of neurotic projection and *akrasia*, the more pedestrian ways that the subject of the RAs influences D.I.R.E., *also* tends to compromise the retributive RAs.

There are already hints of this theme in Sections One and Two of this chapter, since, for example, one of the notable features of Sarah's indignation is that it tends to come to exist, ebbs and flows with the type of violation she reacts to, in accordance with her own past *personal* involvement. She reacts more intensely to lack of generosity than to dishonesty, even to deadly violence, because *she* feels more guilt (whether reasonable or not) over *her own* historical lack of generosity than over these other violations.

The principle of impartiality however, requires the elimination of any reference to the personal or particulars *qua* particulars from moral judgments and responses. Sarah's indignation fails to conform to this principle since her indignation towards Hillary is more intense than her indignation towards the campus murderer *because* it is animated by the unconscious thought that "Hillary's violation is closer in kind *to my own* violation than the plagiarist's or the murderer's." Since the personal reflexive belief is needed to articulate the cause of her intense indignation, a violation of the principle of impartiality has occurred.

But, the point here is this: Strawson underestimates the notion that persons need not resort to the "shady" realm of "unconscious sadism", or Freudian or Nietzschean psychology,

to find similar *breaches of impartiality* in the reactive sphere — because they are *all too common* in *humble*, everyday cases. For instance, it is uncontroversial that our sympathetic feelings and sense of fairness are frequently less intense towards those we believe are different from us. If the object of my moral responses differs from me in some important way, as Hume pointed out, I tend to feel less sympathy and outrage than towards those harmed that are *not like me*. People who are spatially and temporally proximate to the reactor, at least typically, elicit a greater response than those whom are not (Hume, *A Treatise Of*, Book III, Of Morals, Part II, Sec. II, p.p., 484-501, esp. 488). The same is obviously true of other dimensions of proximity and distance, e.g., family membership, race, ethnicity, religion, shared nationalism, and so on. There is a breach to the principle of impartiality where one's response is, at least in part, informed by the thought that "He is like (unlike) *me*, or *my* family, race, religion, nation."

The requirement of impartiality is clearest with what Strawson calls the "impersonal moral" RAs like indignation, where impartiality is built right in. Strawson maintains that indignation is different from a "personal" RA like resentment precisely *because it is* the "impersonal", "disinterested", "vicarious" or "generalized" version of the latter (14). Although indignation and resentment seem similar, if not indistinguishable phenomenological, they are differentiated according to the beliefs they contain. The RA one has towards a violator is simply not indignation unless its object is, at least in part, something like the belief that "S" violated the basic demand for regard to *persons qua persons*. In contrast, the belief embedded in resentment, is more like "S" violated the basic demand for regard towards *me*.

But, notice that this is not to say that the personal RAs are not constrained by the principle of impartiality. Impartiality constrains the personal RAs in a different fashion, namely, as a *moral* requirement rather than also as a *constitutive* belief condition. According to Strawson, one suffers an "abnormal case of moral egocentricity" (15), if for instance, one feels resentment towards R for harming oneself, but fails to feel vicarious resentment (indignation) towards R for harming S in the same way. If Sam feels intense resentment towards Phil for punching him in the nose, but criticizes Judith for "whining about Phil's having punched her in the nose", then something has gone wrong with Sam's resentment, even though that RA is "personal". So, although the object of my resentment does make essential reference to *me qua me*, that does not mean the personal RAs are exempt from the standards of impartiality.

Now, when we bring to the fore these more commonplace breaches of impartiality, it becomes apparent that the "seamy side" of the reactive stance as it actually operates in the realm of inter-personal relationships, and as our attitudes actually reflect our "natural proneness to reactivity" *is* considerably more pervasive than Strawson is willing to grant. Why might Strawson be unwilling to grant that the "seamy side" of the reactive attitudes, including its more humble violations of impartiality, pervades inter-personal relationships and the practice of moral responsibility? This concession raises deep doubts about the prospects for success of Strawson's dialectical *cum* reconciliationist strategy.

Recall that Strawson aims to defuse incompatibilist worries about the "reasonable or appropriate" (6) status of the reactive stance naturalistically. Strawson claims to

> ...fill the lacuna which the pessimist finds in the optimist's account of the concept of moral responsibility, and of the bases of moral condemnation and punishment; and to fill it from the facts as we know them (20).

Rather than going beyond the "facts" Strawson attempts to "fill the lacuna" and reconcile the pessimist with the optimist by appealing to how our RAs *actually* tend to ebb and flow *in the world*. However, avoiding the appeal to metaphysics *vis-a-vis* appealing to facts presents Strawson with a double-edged sword. On the one hand, it gives him the hope of securing the "vital core" of the practice of responsibility in non-controversial facts — free of "panicky libertarian metaphysics". On the other hand, however, it reveals the dialectical vulnerability of any such naturalistic strategy of reconciliation. The "facts" about the actual psychological causes of reactive attitudes, as we discover them, may turn out to be deeply suspect and compromised by such things as ignorance and partiality. Indeed, Strawson's acknowledgement that the reactive stance has its "seamy side" indicates that he exhibits awareness that appealing to the contingencies of the empirical realm brings with it a certain dialectical vulnerability. He admits recent discoveries showing the RAs involve such "seamy" causes as guilt projection and sadism *are* "important" and that we may want to redirect and modify our attitudes in light of these studies. While it is "unlikely" we will ever completely abandon the RAs, he acknowledges that it is not "inconceivable" that we *should* and "the dreams of some philosophers will be realized" (24-25). But, although Strawson thinks that the "seamy side" of the facts make us "rightly mistrustful" of many of the RAs, he does *not* take the "seamy side" seriously enough.

Section 4: Commonplace Breaches of Reactive Impartiality

When we bring to the fore more commonplace breaches of impartiality, it becomes apparent that the "seamy side" of the reactive stance — as it actually operates in the realm of

inter-personal relationships, and as our attitudes actually reflect our "natural proneness to reactivity" — it *is* considerably more pervasive than Strawson seems to grant. For instance, one's own remembered or anticipated *victim hood* can influence the allocation of the RAs. For instance, let us return to campus where Sarah and Hillary learn two violent predators remain on the loose. In addition to a series of brutal sexual assaults on co-eds, there has also been a series of murderous non-sexual assaults on roughly the same number of male students. Suppose that Hillary is outraged about both series of murders and indignant at the police for their inadequate pursuit of the murderers. However, it is apparent from her actions and reactive affect that she is much *more* indignant about the lack of progress in the police investigation of the murderous sexual assaults. For instance, she only attends rallies, which protest the demerits of the one investigation, but not the other. She obviously is much more agitated when talking about the murder of the young women than the young men, and so on.

Here again, we have a case of variation in the D.I.R.E. aspects of a reactor's retributive RA's. This time, rather than an *akratic* break or projection of guilt, Hillary's indignation is, at least in part, explainable in terms of her *greater identification* with the gender of one set of victims than another. The cause of Hillary's more intense and enduring indignation toward police ineptitude in one investigation rather than the other involves her memory of having been sexually assaulted or her own anticipated sense of possible victim hood. But, whether memories of the past or anxieties about the future, the fear that comes from the closer identification with the victim somehow causes Hillary to allocate her indignation in a partial fashion. That is, although Hillary's indignation is purportedly about police incompetence, it varies, in actuality, in accordance with how much *she* has in common with the sex of the victim.

Similarly, race, ethnicity, class and religious affiliation are also common causes of such partial allocations of the retributive RAs. For instance, the Jewish community is typically more indignant about an assault on one of its members than on an assault on a Palestinian. The Palestinian community reciprocates the partiality. Catholics from Northern Ireland are typically more indignant about the death of IRA members than members of the Orange Order. Irish Protestants readily reciprocate the partiality.

Suffice it to say that leaving aside the more exotic forms of the "ambiguous" and "shady" RAs that Strawson acknowledges, e.g., "sadism", "self-deception" and the "projection of guilt" (25), and sticking to more modest versions, does *not* vindicate the retributive RAs since they often result from the reactor's self-reflexive beliefs and attitudes — *contingent* reflexive beliefs and attitudes that are, it seems, rife with *particularity*. The memory or anticipation of one's own suffering or victim hood, one's proximity to those harmed or the perpetrator, and so on — all contingent beliefs and attitudes that breach the principle of impartiality – and seem to explain many instances of the retributive RAs. Thus, when we consider "*all* the facts as we know them," including the more *modest versions* of "seamy" RAs, we see they *are* often grounded in lack of self-knowledge, even self-deception. When we come to recognize that such epistemic flaws ground a retributive RA, then we should give up the RA itself, or some D.I.R.E. aspect of it, at least in that instance. If the occasions are frequent enough, then the reactive stance itself starts to look more and more undesirable.

Notes To Chapter Three

1. *The* essential feature of projection is that the *subject of* the initial painful feeling, i.e., and the *reactor*, is *changed* and becomes the *target* of the other directed reactive attitudes. The subject, that is, "I", in "*I* deserve punishment" becomes changed to "He", as in "*He* deserves punishment." Ironically, the *other* directed reactive attitudes may frequently represent the hope of transforming our neurotic or moral anxiety from our negative *self*-directed reactive attitudes. One who is afraid or guilty about his own lust or neglect of the poor, may obtain some temporary relief for her anxiety by attributing aggressiveness or greed to others, e.g., "*they* have violated the interests of others and deserve punishment — not *I*." Although reactors attempt to eliminate low-self esteem and fear through adopting the retributive stance towards others, scapegoating prevents rather than facilitates, the development of healthy personalities (Hall 89-90).

2. See Alice Miller's notion "narcissistic cathexis" in *Drama of The Gifted Child*, for more on the adverse affects of parental reactive attitudes upon children.

3. See Robert Solomon for more on the notion that emotions based on false beliefs are irrational and breach "epistemic parameters" (Solomon, *The Passions* 381-88).

CHAPTER FOUR

Purifying the Retributive Attitudes?

We have learned, in Chapter Three that when we bring to the fore *more commonplace breaches* of what are, violations of the principle of impartiality, it becomes apparent that what Strawson calls the "seamy side" of the reactive stance, as it actually operates in the realm of inter-personal relationships and as our attitudes actually reflect our "natural proneness to reactivity", *is* considerably more pervasive than Strawson wants to acknowledge. Here, in Chapter Four, I examine *two Strawsonian rejoinders* to my argument that he seriously underestimates the nature, extent, and dialectical importance of the controversial psychological origins of many retributive attitudes.

In Section One, I consider a Strawsonian defense of the retributive attitudes that appeals to a strong role for impartiality ensuring institutions. I argue that although *institutional safeguards* may ensure, to some extent, worries about impartiality, this does *not* help Strawson *dialectically* to convince the libertarian incompatibilist to abandon her metaphysics.

In Section Two, I consider the role the retributive attitudes play in the preservation of *impartial sympathy* and our sense of *common humanity*. I argue that the Strawsonian fares no better here since neglecting the retributive in favor of the objective stance *is* compatible with the preservation of impartial sympathy and deep feelings of compassion towards victims and our common humanity.

Section One: The Institutional Protection of Reactivity from Partiality?

Strawson might object to my worry about the prevalence of partiality and the "seamy" RAs, arguing that this poses no significant problem for him or the practice of moral responsibility. He might argue that worries about dubious causes need not lead us to "be ready to acquiesce to the infliction of injury on offenders in a fashion which we saw to be quite indiscriminate..." and that although accepting the RAs entails "a readiness to acquiesce in the infliction of suffering", this does *not* entail the acceptance of "any punishment for anything deemed an offence" (22, f. 1). If partiality and other dubious factors *do* play a powerful role in commonplace RAs, Strawson might reassure us that *institutions* can preserve impartiality and save the day; "Inside the general structure or web of human attitudes and feelings of which I have been speaking, there is endless room for modification, redirection, criticism, and justification" (23). Strawson might respond to our worries that even *if* the "seamy side" of the reactive attitudes permeates inter-personal relationships, impartiality *can* be restored by such institutional safeguards as the justice system, strong constitutions, and so on. Thus, the Strawsonian might conclude, although it *may* turn out that the *best* explanation for many retributive experiences includes partial (if not the pathological) beliefs and attitudes of the reactor, a more pervasive "seamy side" need not "taint" the practice of moral responsibility itself.

Now, Strawson might be right should he argue that institutional constraints can restore impartiality and cleanse the practice of moral responsibility of any seamy residue left by the retributive RAs. I suggest however, yet again, that this move would seriously blunt the force of Strawson's *reconciliationist strategy*. There is no doubt that in modern societies we *do* rely heavily on formal institutions to ensure the impartiality of the retributive RAs.

For instance, we require judges and jurors to excuse themselves from criminal cases that involve relatives or friends. We look down on nepotism where jobs or favors are gained through one's personal relationship to the powerful. This kind of *impersonality* is the hallmark of *judgments* rendered by modern justice institutions and administration. Is this a problem for Strawson? It seems so: if the judgments of *impersonal* institutions are required to ensure the impartiality of the RAs, then the RAs *themselves* seem to lose their luster. Recall that Strawson must convince the pessimist libertarian to give up his metaphysics:

> The vital thing can be restored by attending to that complicated web of attitudes and feelings which form an *essential* part of moral life as we know it, and which are quite *opposed to the objectivity of attitude*... Because the optimist neglects or misconstrues these ("moral") attitudes, the pessimist rightly claims to find a lacuna in his account. We can fill the lacuna for him. But in return we must demand of the pessimist a surrender of his metaphysics (23, italics mine).

Strawson agrees that "pessimists" libertarians are right to be dissatisfied with the "optimist's" reduction of moral responsibility to social expediency and the neglect of the retributive stance. They are wrong to think, however, that metaphysics is needed to alleviate their "emotional shock" (21) and secure what they want from moral responsibility. As we discussed in Chapter Two, Strawson's reconciliationist strategy hopes to persuade the "pessimist" that they really want metaphysics because they really want that which is missing from the objective attitude, namely, the *non-detached* attitudes and reactions of people directly involved with each other..." (4, italics mine), more specifically, the vitality of the retributive ("moral") attitudes. But, in order to seduce the pessimist into abandoning metaphysics, Strawson must persuade them that the RAs *are* that "vital thing" which is, according to Strawson, equally attractive.

It seems to me, however, that should the "seamy side" of the RAs turn out to require strong institutional safeguards to protect against pervasive violations of the principle of impartiality, then pessimists will have more reason to suspect that the RAs themselves *are not* that "vital thing " missing from the optimist's account: for, it is all too imaginable that a human condition in which *personally experienced* retributive RAs may become *less and less important* to the practice of responsibility in our everyday lives. We saw a version of this problem in R. Jay Wallace's account of moral responsibility. Sober judgment, rather than the actual human experience of the reactive attitudes themselves, played the most important role for moral responsibility. Similarly, if worries about the prejudicial nature of the retributive RAs require strong institutional safeguards, "impersonality" and the judgments of modern administrative institutions seem all too paramount. Strawson suggests that "pessimist" libertarians object to the "optimist" theory of moral responsibility, at least in part, because the latter fails to make the retributive ("moral") attitudes "necessary or appropriate" (20). But if strong institutions are required to ensure the impartiality of the retributive RAs, it gives the pessimist more reason to think that Strawson's theory involves a similar fate: the RAs are *not* that "vital thing" missing from the optimist account.

Section Two: The Sympathy Defense

The Libertarian will be less tempted, so I have argued, to trade in their favorite metaphysical notions about personhood, agent causes and so on, if partiality (whether of modest or seamier sort) taints the retributive RAs. In Section One I argued further that Strawson would be mistaken *dialectically* to defend worries about partiality by appealing to a strong role for institutional safeguards. Although the sober second thought of institutional

judgments may go some distance to ensure against retributive bias, an increased need for impersonality, rules and regulations will leave the libertarian more suspicious about the role reactive feelings play within the actual practice of moral responsibility. In short, *worries about modest but seamy partiality* of the retributive attitudes seem to *taint* the RAs for Strawson's reconciliation project; the pessimist becomes disinterested in reconciliation *for* any increased need for impersonality, rules and regulations will leave the libertarian lacking the vitality and importance he misses under consequentialism. More suspicious about the importance reactive feelings play within the actual practice of moral responsibility, the pessimist becomes *skeptical* about the epistemic and moral prospects of a naturalistic account of moral responsibility, and, the human capacity to adopt the *impartial* or "moral" point of view.

Perhaps I have been too quick, however, to dismiss another Strawsonian defense: What if *impartial* sympathy and *disinterested* concern for our common humanity *typically* ground our other-directed retributive RAs? After all, many of our reactions *do* seem unrelated to the psychological needs of guilt-wracked, fearful or partial selves. Strawson suggests that RAs like resentment and indignation do *not*, at least typically, presuppose the "antecedent personal involvement" (Strawson 17, italics mine) of personal relationships, shared victim hood, race, ethnicity and so on, but rather powerful feelings of "sympathy and of common humanity" (Strawson, "P.F. Strawson Replies" 266, italics mine). According to him, indignation towards others typically ebbs and flows according to our desire that "all" others be "spared suffering", and "the magnitude of the injury". While the "degree to which the agent's will is identified with, or indifferent" to the victim's injury grounds our responses, typically, the connection between how strongly we identify and the magnitude of

their injury is, according to Strawson, "*not* contingent" (21, italics mine). Strawson argues pessimists rightly suffer "emotional shock" in response to the exclusion of the retributive attitudes *for* the attitudes *are* typically "impersonal", "disinterested" (14), *and* reflect deep care and concern for others, not just ourselves *qua* particulars. To exclude the retributive attitudes thus offends humanity since "These practices or attitudes permit, where they do not imply, a certain detachment from the actions or agents, which are their objects" (4). The "impersonal" care and concern of the retributive stance *demands* that, sometimes, we *ought* to feel indignation towards those who inflict harm. The pessimist is right to worry about the optimist's exclusion of retributive feelings *for* this would entail the "generalization of abnormal *egocentricity*" (19, italics mine).

Would Strawson be right to defend the retributive attitudes on the grounds that they presuppose impartial sympathy? It would be pointless to insist that these Strawsonian rejoinders have no merit. Examples of "non-seamy" retributive RAs may indeed be an important part of human experience. Consider our powerful indignation in response to the outrages of war, for instance, or brutality towards children. Perhaps Strawson *is* correct to suggest that such retributive RAs are often based on sympathetic concern for our common humanity, and nothing more suspicious than that.

For instance, in arguing against the Stoic Seneca's sweeping dismissal of the retributive RAs Martha Nussbaum (402), gives a Strawsonian defense of "public anger" citing the powerful example of a large black military officer among the first of the Allied forces to liberate the Auschwitz death camp. Taking one look at the smoldering and wretched remains of the human carnage he spontaneously let out a stream of outraged, indignant charges directed at the Nazis who committed the atrocities. As a child, Elie Wiesel

witnessed the outrage of the officer and thought, "Now, with that anger, humanity has come back" (403). While it may be true, the best explanation for the D.I.R.E. nature of "public anger" includes, at least sometimes, the partial preferences, personal histories, even neurosis of the reactor, Nussbaum points out there are *also* times when the Stoical neglect of the retributive attitudes leads to the neglect of the cultivation of "*humanitas*" and the "detachment" from monstrosities committed against human beings (438). Nussbaum's point is that the kind of angry, indignant response to mass murder exemplified by Elie Wiesel's soldier *is* perfectly understandable from the *emotional* point of view *and* we would be suspicious of anyone who did *not* respond in this fashion:

> On the one hand, (the retributive stance) is closely connected to brutality and a delight in vengeance for its own sake... On the other hand, *not* to (adopt the retributive stance) when horrible things take place seems itself a diminution of one's humanity. In circumstances where evil prevails, (retributive) anger *is an assertion of concern for human well being and human dignity*; and the failure to become angry seems at best "slavish" (as Aristotle put it), at worst a collaboration with evil (403).

Nussbaum expresses the Strawsonian/Aristotelian point that a *failure* to feel retributive emotions, i.e., at the proper time and in the proper degree, indicates a moral shortcoming of some sort, heartlessness towards human suffering for instance, on the part of the observer. Perhaps Nussbaum *et al* are right about this. I do not deny that this is so. The question is just *what* the *moral failing* is, and what it implies about the role of the retributive RAs for the *good life* of flourishing for human beings.

Nussbaum alludes to two ways ethics traditionally answers this question: the failure to adopt the retributive stance against those who commit harms represents 1) the moral failure to uphold justice or 2) the moral failure to respect value. In the first instance, if the

black officer failed to feel profound indignation in the face of a moral abomination like the holocaust he might be failing to live up to the appropriate principle of retributive justice. By failing to be intensely indignant, he fails to make the right judgment about the nature of the wrong committed, and then fails to play his appointed role in meting out the appropriate penalty (429). In the second instance, the absence of indignant affect might be a signal that the officer fails to value the lives and interests of the brutalized victims sufficiently.

Interestingly, Nussbaum seems to stress the latter worry about the Senecan notion that we should purge the retributive attitudes from our lives, not the former. Nussbaum argues that failing to adopt the retributive stance can mean a "selfish sort of non-involvement" towards the victim's suffering and a dubious "egocentricity" (436). Echoing Nussbaum's concern, as we have seen, Strawson thinks the proneness to adopt the retributive attitudes is an inexorable part of our "common humanity" and neglecting them on behalf of others would mean "moral solipsism" (Strawson 14) and the "generalization of abnormal egocentricity" (18).

But, *would* the neglect of retributive attitudes towards those who harm others make society unrecognizably human and engender the "generalization of abnormal egocentricity" as Strawson suggests? *Is* Nussbaum right to worry that extirpation represents the failure to care for victims and a troubling "egocentricity"? It seems to me that defending the retributive RAs on the grounds that neglecting them means the neglect of what makes us recognizably human, the care and concern for our common humanity, seems highly controversial at best: any such defense seems to *assume* that nurturing one's "natural proneness" to the retributive RA like intense indignation *is* the best, even a good way — to nurture one's sense of the deep value of human life. But, as Watson and others have pointed

out, we need *not* choose between "isolation and animosity" (Watson 147). People like Gandhi and Martin Luther King *aspire* to *rid* themselves of the retributive RAs for the retributive attitudes entail the limitation of goodwill and the acquiescence in suffering and punishment. Rather than failing to care sufficiently for the deep value of humanity, Watson reminds us that the repudiation of the retributive sentiments *is* consistent with an "ideal of relationships", and an historically important "ideal of love" (148). Strawson argues the retributive attitudes are not optional but a part of the "framework" of our conception of human society. Watson maintains however, that one could accept Strawson's anthropological thesis about the actual content and practice of holding responsible, but like Gandhi and King still maintain that *abandoning* the retributive attitudes and the practice as it exists is not only conceivable but *desirable*, "for what it expresses itself is destructive of human community" (146).

Moreover, if we ought to be skeptical about the retributive attitude's capacity to respond to such important historical information as a perpetrator's childhood abuse, and, accepting the retributive attitudes entails a readiness to accept the infliction of suffering upon such perpetrators, then it is precisely *because we care deeply for humanity* that we ought to give "pause" (145) to the retributive attitudes. Watson suggests, in other words, that *because* the retributive attitudes *are* unreliable indicators of moral responsibility, 'innocent' persons will (probably) be punished. Therefore, *care and concern* for humanity should involve deep skeptical worries about the retributive stance. The point here is this: it is far from clear that *abandoning* the retributive RAs means *neglecting sympathy* and care for *humanity*. On the contrary, our sense of justice, *sympathy* and care for *humanity* may *require* abandoning the retributive RAs. The Nussbaum neo-Aristotelian critique of the Senecan extirpation of the

retributive RAs falls short: it is far from obvious that nurturing one's proneness to retributive feelings on another's behalf *is* a good way, even the best way, to nurture one's sense of the deep value of human life.

But what if it *could* be shown that Nussbaum and Aristotle were onto something important after all and nurturing one's proneness to indignation on another's behalf *is* a good way, even the best way, to nurture our sense of care and concern for the value of human beings? Even if this controversial premise is true, this does *not* support Strawson's argument that the retributive attitudes are the "vital" core of the *practice of responsibility per se*. Even if the *failure* to feel retributive emotions, i.e., at the proper time and in the proper degree, *does* indicate the failure to feel impartial sympathy for human beings, establishing that such a failure is a *moral* failing would require an additional argument to the effect that it is somehow *unfair* not to feel indignation towards those who violate the basic demand — an argument which neither Strawson nor Nussbaum, nor Wallace seems to provide.

The strength of a Strawsonian "sympathy defense" of the reactive stance also depends upon what kind of *emotion* sympathy is. Recall that first, there is a kind of sympathy which seems almost as controversial, i.e. almost as suspect morally and epistemically as the unconscious projected guilt that seemed rife in the tale of Sarah and Hillary. This kind of sympathy should be called "infantile" or even "egoistic" because it seems like the kind of emotional echo to another's pains adopted by babies, toddlers (perhaps psychopaths(?)), all of whom seemingly lack an understanding of others *as others*, or others as distinct persons. For instance, Baby S hears baby H cry, which prompts Baby S to start crying too. What is going on here? It is not entirely known but it does seem to involve some kind of conditioned response. S hears H cry. S associates crying with *her own pain* and S begins feeling actual

pain (or imagines that she feels pain). In any case, S's "sympathetic" response seems "infantile" or "egoistic" since it does *not* seem to presuppose any thoughts like "that other baby is feeling pain — how terrible — I feel sad for him." After all, babies, like baby S, see everything *as* one thing (an extension of themselves), and have not developed the concept of the *distinctness of persons*. The crying process is pre-conceptual and reflexive, literally. Baby S experiences baby H's pain *as hers*, as the result of associating *her* crying with *her* pain. It is not clear whether babies "learn" this response in any Pavlovian fashion, or whether their "proneness" to associate crying with pain is an instinct. In any case, this kind of sympathy is pre-conceptual and *egoistic*, in the sense that the baby seems totally concerned with no one's pain but *her own* (Blum 509).

The strength of a Strawsonian "sympathy defense" of the reactive stance depends then upon what kind of emotion sympathy is *for* "infantile" or "egoistic" sympathy seems just *as morally and epistemically controversial as* the "*seamier*" RAs, e.g. *akrasia*, projected guilt, lust, and so on, discussed above. In other words, if all sympathy were of the "infantile" variety, then the Strawsonian defense would be in serious trouble — the "seamy side" would constitute all our RAs. I would not hazard to claim anything so radical however, for, as Bishop Butler points out, psychological egoism itself faces trouble (Butler, Sermons 11 and 12). And, there are times when we *do* seem to experience a kind of "adult," or "benevolent" sympathy which *does* seem to involve a real responsiveness to the pains of others — *qua* others — where there *is* cognitive content and little to no confusion about the individuation of persons. Even if we do concede the existence of "adult" varieties of sympathy, however, there is a worrisome strand in Strawson's account of sympathy, which seems to jeopardize his reconciliation strategy. On the one hand, there seems to be the sympathy *for others*,

which seems to cause specific reactive attitudes like indignation. As Strawson insists, we have a natural proneness to feel the "disinterested" (14) retributive RAs *because* we care about the interests of others — *qua* others — as distinct from our own. On the other hand, there is the kind of sympathy or compassion, which *is* available *within the objective* stance. It is important to remember that in Strawson's framework *objectivity* is "emotionally toned in many ways" including, as it does, "pity", a kind of "love" (9) attention to another's "interest" (12) and it does *not* entail coldness *or* lack of concern. After all, the objective stance is not simply that of the inquirer, but that of the Doctor, the healer, the parent and the caregiver. Nussbaum interprets Seneca on a version of sympathy that is more than available from the objective stance.

> The good man is concerned about his fellow citizens in the manner of the Doctor (seeking to) improve the offender... believing that "it is better by far to cure a wrong than to avenge it"... Just as some bodily afflictions respond to a gentle alteration in daily regime, so too the character of some offenders can be treated by "rather gently words" (Nussbaum 417).

The point here is that the *non*-retributive person *can* experience sympathy, "care intensely about humanity, and be motivated on its behalf..." (417). Ridding ourselves of the reactive stance, whether under the pressure of controversial metaphysical *or* psychological worries, need *not*, as Strawson sometimes suggests, result in the absence of impartial sympathy, inter-personal relationships, emotional attachments or care for human beings, let alone the generalization of psychological "egoism". Derk Pereboom is another who agrees that eliminating retributive attitudes *is* compatible with "good inter-personal relationships", "respect" for the "rational capacities of persons", and a mature "love" that makes the aims and desires of another, "one's own" (Pereboom 270-1). It remains to be shown that rich

inter-personal relationships and deep care and concern for human well-being *are* not compatible with the elimination of the retributive feelings. Emotional complexity, warmth and affection seem to fit quite well into the objective stance, particularly when we think of this as the non-retributive stance. Therefore, it seems far from clear how giving up the retributive emotions entails "egoism" or any radical sacrifice of the distinctively and recognizably human, as Strawson sometimes suggests.[1]

It seems then, that even if we *do* experience sympathy beyond the "infantile" or "egoistic" varieties, the retributive attitudes *are* highly problematic nonetheless; defending the retributive RAs on the grounds that they (sometimes) presuppose "impartial" or "adult" sympathy *for others* seems to *assume* that nurturing our disposition for the retributive stance *is* the best, even a good way, to nurture sympathy for others, and moreover, that the sympathy and compassion *for others* available from the objective stance *is* insufficient for human flourishing.

Note For Chapter Four

1. In addition to Derek Pereboom, Galen Strawson and R. Jay Wallace are among those who seem to reject the notion that the experience of the retributive attitudes is necessary for either moral systems or the intense appreciation of the value of human beings.

CHAPTER FIVE

Conclusion: The Dialectical Implications of the Seamy Side

> For Alois Hitler, the suspicion that he might be of Jewish descent was insufferable in the context of the anti-Jewish environment he grew up in. All the plaudits he earned as a customs officer was insufficient to liberate him from the latent rage at the disgrace and humiliation visited on him through no fault of his own. The only thing he could do with impunity was take out his rage on his son Adolf. According to the reports of his daughter Angela, he beat his son mercilessly every day. In an attempt to exorcise his childhood fears, his son nurtured the manic delusion that it was up to him to free not only himself of Jewish blood but also all Germany and later the whole world. Right up to his death in the bunker, Hitler remained a victim of this delusion because all his life his fear of his half-Jewish father had remained locked in his unconscious mind (Miller 1998, 159).

Is Adolf Hitler morally responsible for the consequences of *his* heinous desires? Or, are Adolf Hitler's desires, in some meaningful sense, his father Alois's? Perhaps parental cruelty like that experienced by murderers like Robert Harris *or* Adolf Hitler, "implant" violent desires or engender evil intentions. Unless we develop historical criteria by which the *sources* of desires are identified and assessed, over and above problems raised by the compatibility of moral responsibility with determinism, worries remain about moral responsibility and *justice*. Here, we have emphasized a *different* version of the *source* problem. What *is* the *source* of *our* desire to *hold others* morally responsible?

To sum up thus far, in Chapter One, we discussed the nature of the stand off between "optimists" about moral responsibility, and, their metaphysically attended "pessimist" rivals. We articulated Strawson's *reconciliationist* strategy, according to which, Strawson promises libertarians the reactive attitudes, *the* "vital thing" missing from the optimist's account and a worthy trade for their metaphysics of the self.

In Chapter Two, we explored the log jam over the belief presuppositions of the RAs and articulated a "new" mode of exploring moral responsibility, namely, explanatory potency and inference to the best explanation. However unconsciously, Watson and Wallace initiated the explanatory turn and were right to criticize Strawsonian minimalism for its lack of explanatory power. But, I argued that Watson seemed *too quick* to adopt the skeptical stance about the reactive attitudes, for (as I suggested in the Introduction) a robust historical condition *does* seem open to the compatibilist.

One such historicist compatibilist is R. Jay Wallace. In the effort to define moral responsibility as, at least in part, requiring an historical condition, Wallace faced the incompatibilist generalization arguments head on. We discovered however, his historicist (normative) account did *not* seem to avoid either version of the source problem. For instance, in failing to explain the *seamy source* of important retributive data, Wallace failed to confront and adequately accommodate, among the most important phenomenon which jeopardizes the whole Strawsonian reconciliation strategy, "emotional *akrasia*". I argued phenomena such as *akrasia* dull the luster of what Strawson hopes is the "glittering prize" of the retributive attitudes, for avoiding *akrasia* and other versions of the irrationality of the emotions, leads to a reduced role for the reactive attitudes. Theorists like Wallace make moral *judgment*, rather than the *vivre* of the reactive attitudes, *the* central feature in holding responsible. But, the reduced importance of the reactive attitudes leaves the libertarian uninspired about reconciliation and suspicious about the prospects of a naturalistic account more generally. Like Strawson and Watson, even Wallace failed, in the end, to take the *dialectical implications* of the seamy side seriously enough. Absent their zest and reeking of

dubious origins, the libertarian will remain unconvinced the reactive attitudes present a worthy substitution for the metaphysics of self.

In Chapter Three, I discussed *my* take on the further explanatory inadequacies of Strawsonian minimalism and presented the hypothesis that many, if not all, RAs presuppose what amount to dubious *epistemic* and *moral* beliefs and attitudes about *our own* history and interests. O*ver and above* any beliefs about others, I argued on many occasions important aspects of our RAs are *best explained* in terms of the "seamy side". We saw one way the contingent reflexive beliefs of the self influence the D.I.R.E. of retributive attitudes towards others is, the projection of guilt. We found this version of the seamy side (like most versions) problematic because the reactor believes his or her retributive attitudes are *solely* about others. But, to the extent the retributive attitudes involve the projection of guilt, they are *not* solely about others. Therefore, the retributive attitudes that involve the projection of guilt involve false beliefs and *are* criticizable from the standpoint of theoretical rationality. I argued further over and above neurosis and pathology, the "seamy side" of the retributive attitudes also involves more *modest versions*. Reflexive beliefs and attitudes involving proximity to the perpetrator, the person harmed, shared history, ethnicity, religion, victim hood and so on, frequently provided a *better explanation* than Strawsonian minimalism alone for retributive experience. Although such "modest" retributive attitudes appeared, on the face of it, less worrisome than their pathological counterparts, they were "seamy" nonetheless; both presuppose *dubious moral and epistemic beliefs* involving *lapses of impartiality*, if not *rationality*.

In Chapter Four, we examined two objections to my argument that he seriously underestimates the nature, extent, and *dialectical import* of the RA's "seamy side". I

considered a Strawsonian defense of the retributive attitudes that appealed to a strong role for impartiality ensuring *institutions*. But, we saw the sober second thought of institutional judgments, increased impersonality, rules and regulations would leave the libertarian uninspired and more suspicious about the reducible role reactive feelings play within the *actual* practice of moral responsibility. While institutional safeguards will sooth some anxieties, such detached impersonality, it fails to vindicate the importance of the retributive attitudes themselves. We considered another Strawsonian rejoinder that faired no better. Appealing to the importance of *impartial sympathy* or the care and concern for *common humanity*, seemed to *assume too much*, for deep feelings of care and sympathy for human beings *do* seem compatible with adopting the objective stance.

In short, I have argued that the *source* of the retributive RAs and the *desire to hold others* morally responsible is frequently dubious. Historical *sources* such as partiality and neurosis — *within the reactor* — explain, at least in part, the nature of responses to the *sources* of other's desires. Rather than *impartial care* for *all* human beings, the desire to adopt the retributive stance and hold others morally responsible, e.g., Robert Harris, Bill Clinton and Muslim highjackers, may all too frequently involve sources in guilt projection, proximity to the victim, ego aggrandizement, and so on. If such morally suspect sources frequently ground the *reactors desire*, then, morally sensitive responsiveness to the *sources of other's desires*, e.g., child abuse, addiction, desperation and so on, are all *too* rare. In addition to the compatibility of moral responsibility and determinism, the pervasiveness of the reactor's self, or the "seamy side" of the retributive attitudes, raises worries about the compatibility of moral responsibility and *justice*.

But, what might this more pervasive but modest version of the "seamy side" of reactivity mean for Strawson's reconciliation project more specifically? *Worries about the modest but seamy partiality* of the retributive attitudes *taint* the RAs for, in addition to general worries about *justice,* we also discovered a more complete spectrum of "seamy" retributive attitudes raises concerns about *lack of self-knowledge* or epistemic irrationality. We saw that beliefs and attitudes involving prior victim hood, ethnicity, lust, and so on, may provide a *better explanation* than Strawsonian minimalism alone for the D.I.R.E. nature of our RAs. If the incompatibilist *can* be convinced that the reactive stance does *not* presuppose robust metaphysical beliefs, worries about the pervasiveness of the "seamy side of the facts" should, nevertheless, give the incompatibilist significant *dialectical* pause. After all, as we have seen, Strawson seriously *underestimates the controversial nature* of the beliefs and attitudes, which the retributive RAs often involve, for, the RAs *frequently* commit the reactive person to psychological, if not metaphysical, beliefs that *violate principles of impartiality*, if not *epistemic rationality*. This *is* important in the *dialectical context* of "Freedom and Resentment," because Strawson wishes to *reconcile* the compatibilists and incompatibilists by persuading libertarian pessimists to give up their "panicky metaphysics" (25) of free will, agent-causes, noumenal selves and the like, in return for a concession from optimists that place exclusive emphasis on the *consequentialist* aspects of moral responsibility.

But, if the lines of argument I have been pursuing are plausible, they certainly do *not* provide a "knock down" refutation of any of Strawson's central *compatibilist theses*. I have only defended the weaker thesis, namely, that some (not all) RAs are "tarnished" by their origins, and the point then, is not, that all RAs must be purged, but rather, that the proneness

of RAs to "tarnishment" jeopardizes Strawson's strategy of reconciling libertarians to determinism. I claim that even this weaker thesis represents a considerable challenge to his *reconciliationist project*. Recall that Strawson's crucial device is to offer the incompatibilist or pessimist a trade: he is to give up his metaphysics of the acting self in return for a *glittering* prize. As we have seen, Strawson attempts to seduce the pessimist in part by arguing that our retributive attitudes are *not* grounded in *expediency*, at least typically, but represent that which *is* distinctly and recognizably human (21). This incentive will be some *natural* aspect of the practice of responsibility. We will find it in "the facts as we know them". It will *fill the breach* in the practice left so gaping wide by mere consequentialist compatibilists. It will be an *extraordinarily attractive* aspect of that practice quite on its own. It is the set of *emotions* we feel when we take up the *reactive stance* (1).

Strawson can afford to tolerate the suggestion that this glittering prize has its tarnished spots here and there. However, he cannot hope to win over the incompatibilist if the tarnish is spread over too much of its surface. In this thesis, especially in Chapters Three and Four, I have argued that the tarnish *has* spread further than Strawson would like. It is very well for him to downplay the seriousness of his dialectical problem by focusing on the dark and unconscious aspects of the tarnish or "seamy side". This minimizes the problem, however, making it *too easy* for him to claim that such Freudian and Nietzschean explanations for reactivity are *rare* and highly *exaggerated.* I myself may have been guilty of exaggerating their role in Section Two, Chapter Three. However, I have tried to redress the balance in Chapter Three, Sections Three and Four, by stressing the more modest or *commonplace* aspects of the seamy side, i.e., all those *everyday* breaches of impartiality, which tend to make for a satisfactory explanation of D.I.R.E. and the retributive RAs.

Impartiality and the lack of self-knowledge thus thwarts Strawson's dialectical needs, for Strawson must persuade the pessimist the "internal" point of view and the reactive attitudes *can* readily provide "*all*" we mean by morality. The Strawsonian argument for the sufficiency of the "internal" point of view includes the notion that, at least typically, retributive attitudes *are* moral and do *not* breach the principle of impartiality. Strawson maintains that it is this "impersonal" or "disinterested" point of view that entitles the retributive emotions to the qualification "moral" (Strawson 14). But, Strawson's argument also involves the idea that the reactive attitudes are a permanent part of human nature and are, at least to a large extent, immune from rational control. In order to persuade the pessimist to abandon her search for external justifications and embrace the internal point of view, she must be convinced then, that human nature *itself* "controls" the moral content of the retributive attitudes. But, should the pessimist *suspect* that many retributive attitudes presuppose contingent beliefs and attitudes about the self, e.g., involving the reactor's ethnic origins, familial attachments, shared victim-hood, and so on, then Strawson will fail to persuade the pessimist human nature is impartial or "controls" the *moral* content of the retributive attitudes. I have argued that the *best* explanation for human retributive experience *will* aggravate these suspicions. If the modest "seamy side" does provoke such suspicions, then the pessimist will fail to be assured that the reactive attitudes and the "internal" point of view *can* readily provide "*all*" we mean by morality.[1]

Beyond any worries posed by incompatibilism, skepticism about the nature and causes of the retributive attitudes gives the libertarian incompatibilist more reason to resist Strawson's offer to trade metaphysics for the RAs. Over and above concerns over metaphysics, Strawson's reconciliationist project is dialectically fragile, if not fatally flawed.

When the incompatibilist considers "all the facts as we know them," i.e., dark *and* modest versions of "seamy" reactive attitudes, the RAs loose their luster: the existence of unacknowledged *reflexive* beliefs and attitudes about a *plurality* of possible factors, such as the person's own *moral guilt*, her *ethnic* and *familial* attachments, or her own *role as a victim* in similar situations in the past play a critical role in explaining his or her retributive RAs towards *others*. Consequently, "moral" attitudes we adopt *towards others* frequently involve certain *beliefs about the reactor herself, and* many of the retributive RAs *are* grounded on the false "meta-belief" that they are merely about and merely caused by beliefs about the histories and intentions of others. Insofar as the reactor is *ignorant* of the real grounds of his or her own retributive attitudes, the reactor *lacks self-knowledge*. Insofar as *lack of self-knowledge* is a defect, the RAs sustained by such a lack are themselves defective in that they are grounded in false beliefs about the real grounds of one's own RAs. Therefore, since the genesis of so many other-directed RAs involves self deception, or at the least, a failure of self-knowledge, Strawson's *reconciliationist* project of demonstrating that the RAs are generally free of questionable facts *is* deeply problematical.

Contrary to what Strawson might think, the "seamy side" of the RAs *does* present problems for reconciling the disputants in the moral responsibility debate — and the dark side need *not* even be typical to *taint the RAs*. Tainted goods tend to make one lose one's appetite. Tainted RAs will make the libertarian incompatibilist lose the appetite for reconciliation. Concerns about the *source* of the *reactor's* desires to adopt the reactive attitudes towards others and attribute moral responsibility may pose then, as vexing a problem for the vindication of the reactive attitudes as responsiveness to historical factors relevant to the source of *other's* desires and actions. Worries that the *proneness* to adopt the

reactive stance is all *too often* grounded in human *ignorance* and *partiality*, will leave the pessimist unwilling to abandon her metaphysics, and, unwilling to reconcile with the optimists in the moral responsibility debate. Contrary to what Strawson might think, the "seamy side" of the reactive stance *does* leave the pessimist unconvinced the proneness to adopt the reactive stance is *all* that is meant by morality, or *the* "vital thing" missing from the optimist account. Philosophers do tend to *underestimate* the dangers and dialectical implications of the "seamy side" of the reactive stance, *not* exaggerate them, as Strawson suggests.

I have been developing a kind of argument by attrition, designed to take enough of the glitter off the reactive stance that it will no longer be a very tempting substitute for the libertarian's favorite metaphysics. It goes without saying that all parties to the dispute are positioned to judge the general success or failure of my criticisms of Strawson's reconciliationist project, not to mention the success or failure of his project itself. However, in the dialectical context of this thesis, only the libertarian will have a *dialectical interest* in judging whether the luster has been taken off the retributive RAs by the considerations I advance. Schlickian compatibilists need not resist the gist of Strawson's project nearly as much as incompatibilists (of both types) for Schlickian compatibilists *do not* have as much at stake. That is, they have much less dialectical difficulty in making their "concession" and conceding the crucial role the retributive attitudes play in the practice of moral responsibility than the determinist incompatibilists do in withdrawing their general incompatibility claim, or the libertarians do in abandoning their metaphysics of the self. I conclude, stressing the libertarian reactions, because they are the direct target of Strawson's attempt to secure a

"withdrawal" of what he takes to be their "panicky" metaphysics of the self. Only the libertarian, then, can fully judge the extent of my success.

Note For Chapter Five

1. As discussed, Strawson maintains the descriptive thesis arguing ordinary retributive attitudes do not, as a matter of anthropological fact, respond to beliefs about determinism. But, persuading pessimists to abandon metaphysics and commit to the retributive stance also requires adopting the prescriptive thesis, namely, the way ordinary retributive attitudes are, *is* the way they *ought* to be. Is Strawson's reconciliationist project successful? Can the pessimist correctly infer anything about how the RAs *ought* to respond from how the RAs *do* respond? Strawson seeks to avoid the *non sequitur* by appealing to the notion of "internal" justification: "questions of justification are internal to the structure or relate to modifications internal to it" (23). As we have seen, for Strawson, the *appropriateness* of the reactive attitudes is derived from "internal" considerations, namely, the profound degree of *care and concern* we adopt *about the intentions* of others towards, not only ourselves, but also others. On the other hand, we do *not* adopt a profound degree of care and concern about metaphysical determinism. In other words, Strawson hopes to ground the *appropriateness* of the reactive attitudes, in the profound degree of care and concern for *all humanity* they presuppose. There is no need to "over-intellectualize" or justify the reactive attitudes "externally", *for* the practice *is* grounded in our internal, life giving commitment to them. So, for Strawson, this commitment, this human sense of appropriateness *is* provided by the human proneness to adopt the reactive stance, itself. But, Strawson must persuade the pessimist this internal point of view and the proneness to adopt the reactive stance *is sufficient* for morality. Since, as we have seen, the RAs are an inescapable part of

humanity, and are, to a large extent immune from rational control, this means, Strawson must persuade the pessimist the *proneness* to adopt the reactive attitudes *is* the *proneness* to adopt the *moral point of view*. But, if, as I have argued, deep worries remain that human proneness to adopt the reactive stance *is not* the proneness to adopt the moral point of view, the pessimist *will* remain anxious about the sufficiency of the internal point of view — and the reactive attitudes. Full of anxiety, the pessimist will cling to his metaphysics and the hope of *external* justification.

WORKS CITED

Alcoholics Anonymous. *Alcoholics Anonymous: The Story of How Many Thousands of Men and Women Have Recovered from Alcoholism*. New York: AA World Services, 1976.

Aristotle. *Nicomachean Ethics*. Tr. Terence Irwin. Indianapolis: Hackett Publishing, 1985.

Bennett, Jonathan. "Accountability." *Philosophical Subjects*. Ed. Zak Van Straatan. (1980): 14-47.

Benson, Paul. "The Moral Importance of Free Action." *Southern Journal of Philosophy*. 28 (1) (1990): 1-18.

Beroksky, Bernard. Ed. *Free Will and Determinism*. New York: Harper & Row, 1966.

Bilgrami, Akeel. "Self-Knowledge and Resentment." *Knowing Our Own Minds*. Wright, Crispin, et al. Oxford: Clarendon Press, 1998.

Blum, Lawrence. "Compassion," *Explaining Emotions*. Ed. Amelie Oksenberg Rorty. Berkeley: University of California Press, 1980. 507-517.

Brown, Eric. "Sympathy and Moral Objectivity." *American Philosophical Quarterly* 23 (2) (April 1986): 179-187.

Butler, Joseph. *Fifteen Sermons Preached At The Rolls Chapel*. (1726). Ed. J.H. Bernard. London: SPCK, 1970.

Damasio, Antonia. *Descartes' Error: Emotion, Reason and the Human Brain*. New York: G.P. Putnam, 1994.

Downie, R. S. "Objective and Reactive Attitudes." *Analysis* 27 (1966): 33-39.

Elster, John. *Strong Feelings: Emotions, Addiction, and Human Behavior*. Cambridge: The MIT Press, 1999.

Fischer, John M. and Mark Ravizza, Eds. *Perspectives on Moral Responsibility*. Ithaca: Cornell University Press, 1993.

Flanagan, Owen. *Varieties of Moral Responsibility: Ethics and Psychological Realism*. Cambridge: Harvard University Press, 1993.

Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66 (1969): 829-839.

_____. "Freedom of the Will and the Concept of the Reason." *Journal of Philosophy* 68 (1971): 5-20.

Hahn, Lewis. Ed. *The Philosophy of P.F. Strawson*. Chicago: Open-Court, 1980.

Hall, Calvin S. *A Primer of Freudian Psychology*. New York: Penguin-Meridian, 1999.

Hobbes, Thomas. *Leviathan*. (1651) Ed. Richard Tuck New York: Cambridge University Press, 1991.

Honderich, Ted. Ed. *The Oxford Companion to Philosophy*. Oxford: Oxford University Press, 1995.

_____. *A Theory of Determinism: The Mind, Neuroscience, and Life Hopes*, New York: Oxford University Press, 1988.

Hospers, John. "Meaning and Free Will." *Philosophy and Phenomenological Research* 10 (March 1950): 307-330.

_____. "What Means This Freedom?" Ed. B. Berofsky. *Free Will and Determinism*. New York: Harper and Row, 1966, 26-45.

Hume, David. *A Treatise of Human Nature*. (1739-40). Ed. P. H. Nidditch. Oxford: Clarendon Press, 1978.

_____. *An Enquiry Concerning Human Understanding*. (1748). Ed. T. L. Beauchamp. Oxford: Oxford University Press, 1999.

Kane, Robert H. *The Significance of Free Will*. New York: Oxford University Press, 1998.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. (1785). Tr. H.J. Paton. New York: Harper and Row, 1964.

Kennedy, Duncan. "Freedom and Constraint in Adjudication: A Critical Phenomenology." Eds. Frederick Schauer and Walter Sinnott-Armstrong. *The Philosophy of Law: Classic and Contemporary Readings with Commentary*. Fort Worth: Harcourt Brace, 1996, 62-69.

Lear, Jonathan. *Happiness, Death, and the Remainder of Life*. Cambridge, Mass.: Harvard University Press, 2000.

Locke, John. *An Essay Concerning Human Understanding*. (1689). Ed. Peter H. Nidditch. Oxford: Clarendon Press, 1975.

Lloyd, Genevieve. "Maleness, Metaphor, and the 'Crisis' of Reason." Ed. Diana Tietjens Meyers. *Feminist Social Thought: A Reader*. New York: Routledge, 1997, 286-302.

Marx, Karl. *Capital: A Critique of Political Economy*. Ed. Frederick Engels. New York: International Publishers, 1967.

Mate, Gabor. *Scattered Minds: A New Look at the Origins and Healing of Attention Deficit Disorder*. Toronto: Alfred A. Knopf Canada, 1999.

Miller, Alice. *The Drama of the Gifted Child*. Translated from the German by Ruth Ward. New York: Basic Books, 1981.

_____. *Paths of Life: Seven Scenarios*. Translated from the German by Andrew Jackson. New York: Vintage Books, 1999.

Miri, Mrinal. "On Knowing Another Person." *The Journal of Value Inquiry* 18 (1984): 3-12.

McKenna, Michael S. "The Limits of Evil and the Role of Moral Address: A Defense of Strawsonian Compatibilism." *Journal of Ethics* 2.2 (1998): 123-142.

Nagel, Thomas. *The View from Nowhere*. New York: Oxford University Press, 1986.

Nietzsche, Friedrich. *Beyond Good and Evil*. Ed. W. Kaufmann. New York: Vintage, 1966.

_____. *The Genealogy of Morals*. Tr. D. Smith. New York: Oxford University Press, 1997.

Nussbaum, Martha. *The Therapy of Desire: Theory and Practice in Hellenistic Ethics*. Princeton: Princeton University Press, 1994.

Otuska, Michael. "Incompatibilism and the Avoidability of Blame." *Ethics* 108 (July, 1998): 685-701.

Paul, Ellen F., Fred D. Miller and Jeffrey Paul, Eds. *Responsibility*. Cambridge: Cambridge University Press, 1999.

Pears, David. "Strawson on Freedom and Resentment." Ed. Lewis E. Hahn. *The Philosophy of P.F. Strawson*. Chicago: Open-Court, 1980.

Pereboom, Derk. "Determinism al Dente." Ed. D. Pereboom. *Free Will*. Indianapolis: Hackett Publishing Co., 1997: 242-272.

Riker, John H. *Ethics and the Discovery of the Unconscious*. Albany: SUNY Press, 1997.

Russell, Paul. "Strawson's Way of Naturalizing Responsibility." *Ethics* 102 (1992): 287-302.

Schlick, Morris. "When is a Man Responsible?" Ed. B. Beroksky. *Free Will and Determinism*. New York: Harper & Row, 1966.

Sen, Pranab Kumar. *The Philosophy of P.F. Strawson*. New Delhi: Indian-Coun-Phil-Res, 1995.

Seneca, Lucius Annaeus. *Seneca*. Baltimore: Johns Hopkins University Press, 1992.

Shoeman, Ferdinand, Ed. *Responsibility, Character and the Emotion*. Cambridge: Cambridge University Press, 1987.

Singer, Peter, Ed. A *Companion to Ethics*. Cambridge, Mass.: Blackwell Reference, 1991.

Solomon, Robert C. *The Passions: The Myth of Human Emotion*. New York: Anchor Press/Doubleday, 1976.

_____. "On Emotions as Judgments." *American Philosophical Quarterly* 25 (2) (April, 1988): 183-191.

_____. *About Love: Reinventing Romance for Our Times*. New York: Simon & Shuster, 1998.

Strawson, Galen. *Freedom and Belief*. Oxford: Clarendon Press, 1986.

_____. "On Freedom and Resentment." Eds. John M. Fischer and Mark Ravizza. *Perspectives*. (1993): 67-101.

_____. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75 (1-2) (1994): 5-24.

Strawson, P.F. "Replies to Ayer and Bennett." *The Philosophy of P.F. Strawson*. Ed. L.Hahn. Chicago: Open-Court. (1998): 260-266.

_____. "Freedom and Resentment." in *Freedom, Resentment and Other Essays*. London: Methuen, (1974): 1-25.

Van Inwagen, Peter. "Ability and Responsibility." *Philosophical Review* 87 (1978): 201-224.

Van Straatan, Zak, Ed. *Philosophical Subjects: Essays in Honour of P.F. Strawson*. Oxford: Clarendon, 1980.

Wallace, R. Jay. *Responsibility and the Moral Sentiments*. Cambridge, Mass.: Harvard University Press, 1996.

Watson, Gary. Ed. *Free Will*. New York: Oxford University Press, 1982.

_____. "Free Agency." *Journal of Philosophy* 72 (1975): 205-20. Rpt. in Watson. (1982).

_____. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." Rpt. in *Perspectives on Moral Responsibility*. Eds. John M. Fischer, and Mark Ravizza. Ithaca: Cornell University Press, 1993: 119-151. Also in F. Shoeman, *Responsibility*: 256-286.

Weihofen, Henry. *The Urge to Punish*. New York: Farrar, Straus & Cudahy, 1956.

Williams, Bernard. *Shame and Necessity*. Berkeley: University of California Press, 1993.

Wooton, B. *Crime and the Criminal Law*. London: Steven & Sons, 1981.

Zimmerman, David. "Hierarchical Motivation and Freedom of the Will." *Pacific Philosophical Quarterly* 62 (1981): 354-368.

Zimmerman, David. "Thinking with Your Hypothalamus: Reflections on a Cognitive Role for the Reactive Emotions." *Philosophy and Phenomenological Research* (forthcoming).