7-26-2012

Moral Responsibility "Expressivism," Luck, and Revision

Kyle Walker
Georgia State University

MORAL RESPONSIBILITY "EXPRESSIVISM," LUCK, AND REVISION

by

KYLE L. WALKER

Under the Direction of Dr. Eddy Nahmias

ABSTRACT

In his 1962 paper "Freedom and Resentment," Peter Strawson attempts to reconcile incompatibilism and compatibilism about moral responsibility and determinism. First, I present the error committed by the proponents of both these traditional views, which Strawson diagnoses as the source of their standoff, and the remedy Strawson offers to avoid the conflict. Second, I reconstruct the two arguments Strawson offers for a theory of moral responsibility that is based on his proposed remedy. Third, I present and respond to two proposed problems for the Strawsonian theory: moral luck and revisionism. I conclude with a summary of my defense of Strawsonian "expressivism" about moral responsibility, and offer suggestions for further research.

INDEX WORDS: Moral responsibility, Moral luck, P. F. Strawson, Revisionism, Expressivism, Meta-ethics, Moral sentiments, Reactive attitudes, Moral psychology

MORAL RESPONSIBILITY "EXPRESSIVISM," LUCK, AND REVISION

by

KYLE L. WALKER

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in the College of Arts and Sciences

Georgia State University

2012

MORAL RESPONSIBILITY "EXPRESSIVISM," LUCK, AND REVISION

by

KYLE L. WALKER

Committee Chair: Eddy Nahmias

Committee: Andrew I. Cohen

Daniel Weiskopf

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2012

ACKNOWLEDGEMENTS

Many people have given helpful comments on previous versions of this thesis. I'd like to thank in particular J.P. Messina, Bill Glod, Dave Schmidtz, Michael McKenna, and the participants at IHS's Scholarship and a Free Society, 2012. For their companionship at Georgia State and thought provoking conversation that contributed to the ideas presented here, I would also like to thank Brad Wissmueller, Anaïs Stenson, Rueben Stern, Kyle Hirsch, Ben Fischer, Cindy Phillips, Cléo Grimaldi, Mike Huddleson, Hunter Thomsen, Marcos Gonzalez, Vincent Abruzzo, and Carson Young. In addition I would like to thank the members of the Moral Psychology Reading Group, especially Jason Shepard, Sam Sims, Shawn Murray, James Digiovanni, Noel Martin, and Getty Lustila. I am also indebted to my committee members Dr. Andrew I. Cohen, Dr. Daniel Weiskopf, and especially my thesis director, Dr. Eddy Nahmias. My parents, Mary and Randy Walker, and Marie have continuously been supportive throughout the writing of this thesis, and they also deserve special thanks.

TABLE OF CONTENTS

1. INTRODUCTION

In his 1962 paper “Freedom and Resentment," Peter Strawson attempts to reconcile incompatibilism and compatibilism about moral responsibility and determinism. First, I present the error committed by the proponents of both these traditional views, which Strawson diagnoses as the source of their standoff, and the remedy Strawson offers to avoid the conflict. Second, I reconstruct the two arguments Strawson offers for a theory of moral responsibility that is based on his proposed remedy. Third, I present and respond to two proposed problems for the Strawsonian theory: moral luck and revisionism. I conclude with a summary of my defense of Strawsonian “expressivism” about moral responsibility, and offer suggestions for further research.

It is worth noting upfront that my reconstruction of Strawson’s arguments offers an interpretation of his view that is outside of the mainstream. The mainstream interpretation holds that Strawson is a meta-ethical expressivist about moral responsibility judgments in the traditional non-cognitivist sense (see Watson (1987) and Vargas (2005, p. 406)). A *non-cognitivist* about some set of judgments holds that such judgments have no truth value. In this sense, expressivism holds that moral judgments are nothing more than the *expression* of the preferences of the judging agent. Strawson does argue that intuitive responses to another person’s good or ill will – what he calls “reactive attitudes” or “moral emotions” – are the basis for his outline of a theory of moral responsibility. Yet, in section 3.1, I interpret Strawson instead as a *cognitivist* about moral responsibility judgments – one who thinks that such judgments are either true or false. Interpreted as a cognitivist expressivist, I attempt to show that a Strawsonian is much better equipped to respond to the objections from luck and revision.

2. STRAWSONIAN "EXPRESSIVISM" ABOUT MORAL RESPONSIBILITY

Strawson's aim in formulating his theory of moral responsibility is to dissolve a dispute between two opposing positions regarding the problem of free will and determinism. The first position, *incompatibilism*, holds that the truth of determinism would universally undermine free will, and for this reason Strawson calls it the pessimistic view. The second position, *compatibilism,* holds that the truth of determinism would not undermine free will, and Strawson correspondingly refers to it as the optimistic view. The practical import of this debate lies in the justification of everyday punitive or approbative practices, such as praise and blame or reward and punishment.[1] Both pessimists and optimists accept that free will is required for the justification of these practices. As a consequence, a pessimist should think that if determinism is true, these practices ought to be abandoned, while an optimist should think there remains the possibility for an acceptable justification of these practices even if determinism is true. Strawson attempts to refocus this debate away from free will and towards moral responsibility, and specifically the attitudes and practices involved in holding responsible. After diagnosing the source of the traditional dispute, arguing that it results from both parties adopting an "objective stance" that ignores everyday attitudes and practices, Strawson suggests that we adopt the "participant stance," which proceeds first by description and analysis of the attitudes and practices involved in holding each other responsible. By adopting this stance, Strawson argues that we can avoid the traditional stalemate.

[1] In this paper I do not address the issue of state institutionalized reward and punishment, and there is no indication that Strawson is concerned with such practices either. I am concerned with punitive and approbative practices among individuals, such as thanks and reproach or social inclusion and ostracization. A theory of state and other forms of institutionalized punitive and approbative practice requires considerations that do not arise for a theory of responsibility among individuals. For example, it is important to consider the coercive power of the state, the fact that it extracts taxes from the very people on whom it exerts punishment, and the fact that institutions generally do not interact in the interpersonal human manner that will become important for Strawson's theory.

2.1 The Objective Stance

Strawson exposes the "objective stance" as the source of error that results in the standoff between the traditional views pertaining to the relation between free will and determinism.[2] For the most part, both sides in this debate are prepared to accept that human beings are, at least sometimes, free and morally responsible agents. On the one hand, pessimists will have to argue that determinism is false, so that the necessary condition for moral responsibility is fulfilled. On the other hand, optimists will usually argue that fulfillment of some other condition is sufficient for moral responsibility. For the optimists to whom Strawson addresses his argument, the efficacy of punitive and approbative practices in regulating social behavior is sufficient justification for holding others morally responsible even if determinism is true. I will call these optimists *classical optimists* to distinguish their view from the brand of optimism developed by Strawson. Following Watson (1987), I will refer to the Strawsonian position as *expressivism*.

It is significant that the debate between pessimists and classical optimists revolves around the metaphysical thesis of determinism. Strawson explains that this is the consequence of an error committed by both parties. The pessimist worries that no one would ever really *deserve* praise or blame if determinism is true, and appeals too quickly to the "panicky metaphysics" (1962, p. 203) of agent causation to alleviate this worry – insisting that agents can somehow break into otherwise deterministic causal chains. The classical optimist answers this worry by dismissing the pessimist's sense of desert and justifying responsibility attribution by appealing to the calculated efficacy of our practices in regulating social behavior. For Strawson, both parties, while accepting among the "facts as we know them" that people are capable of morally

[2] Nothing in Strawson's paper turns on having a precise conception of determinism, though it is important that however it is understood, it is, for incompatibilists, a thesis that would *universally* rule out free will. In fact, he specifically tries to show that his reconciliation should work without even knowing what the thesis of determinism means (p. 191).

responsible agency, go beyond the facts in an attempt to justify our punitive and approbative practices by appeal to objective criteria that are external to the practices themselves. Thus, Strawson diagnoses these disputants with "over-intellectualizing" the facts as follows:

> The optimist's style of over-intellectualizing the facts is that of a characteristically incomplete empiricism, a one-eyed utilitarianism. He seeks to find an adequate basis for certain social practices in calculated consequences, and loses sight (perhaps wishes to lose sight) of the human attitudes of which these practices are, in part, the expression. The pessimist does not lose sight of these attitudes, but is unable to accept the fact that it is just these attitudes themselves which fill the gap in the optimist's account. (1962, p.201)

Classical optimists and pessimists both seek external justification, i.e. justification which overcomes the challenge posed by determinism, for practices that involve holding each other responsible. For classical optimists, any such practice is justified when, after tally of all the relevant positive and negative consequences of that practice, it is determined that the benefits for social cohesion outweigh the costs. For pessimists, holding someone responsible is justified if that individual satisfies the requirements of agent causation. Appeal to these conditions derives from the fact that both parties seek justification for our practices from an impossible, or at least unnecessary, "objective" standpoint, which is external to the practices themselves. Only an analysis of moral responsibility from this perspective leads to determinism becoming a central problem, and allows agent causation and the cold calculus of consequentialism to count as appropriate responses to this problem. Furthermore, the objective stance gets the reasons wrong for holding others responsible in practice. No one, at the time of offense or benefit, thinks they are justified in blaming or praising because the object of their attitude has agent causal powers or because they have calculated all of the relevant consequences.[3] Thus, the objective stance explains the standoff reached in the debate surrounding determinism. Strawson argues in favor of

[3] Thanks to Bill Glod for this helpful point.

an alternative stance by which we can circumvent this standoff and better understand the reasons for holding responsible.

2.2 The Participant Stance

The alternative proposed by Strawson is to allow another set of "facts as we know them" – namely, the human attitudes, and especially the "moral sentiments" – to take their rightful place in a proper understanding of moral responsibility and its associated practices. Strawson insists on an assumption, which he refers to as "the central commonplace," that this view rests upon, namely, "the very great importance that we attach to the attitudes and intentions toward us *of* other human beings, and the great extent to which our personal feelings and reactions depend upon, or involve, our beliefs about these attitudes and intentions" (1962, p. 186). Among the human attitudes, the *reactive attitudes* in particular arise in response to the good or ill will of one person toward another person, which we encounter in ordinary interpersonal interactions. On this view, reactive attitudes constitute praise or blame and *express* the interpersonal demands that we place upon one another – hence the designation *expressivism* about moral responsibility. These attitudes are further divided into three sub-types: personal reactive attitudes, vicarious reactive attitudes, and self-reactive attitudes.

The sub-types of reactive attitudes are differentiated by three classes of objects in response to which they respectively arise. Personal reactive attitudes arise in response to the attitude expressed toward oneself by another person, and among these Strawson counts, on the negative side, resentment, and on the positive side, gratitude. Vicarious reactive attitudes arise in response to the good or ill will of other people toward still other members of what now emerges as an interpersonal moral community. Among this second sub-type are included, for example,

moral disapprobation or indignation and moral approval. Finally, self-reactive attitudes arise within oneself in response to the quality of one's own will toward other members of the moral community; examples include guilt, pride, and the sense of obligation. Defect or excess – and often mere adequacy – in meeting the demands we place on each other, and on ourselves, stimulate the reactive attitudes. These attitudes are implicated in responsibility precisely because disposition to react with these attitudes is constituted by interpersonal demands that we expect members of the moral community to respect.

Pessimists and classical optimists overlook the reactive attitudes; and, according to Strawson, reconciliation between these disputants requires recognizing that reactive attitudes are expressions of the demands we place upon one another. In particular, three consequences of this recognition can move us toward the reconciliation Strawson seeks. First, reactive attitudes answer the pessimistic worry that consequentialist justifications of moral responsibility leave out "something vital," without the usual, less parsimonious, pessimist appeal to agent causation. Second, description and analysis of these attitudes provide an adequate theory of responsibility, so determinism is rendered irrelevant. Third, the external justification for our practices sought by pessimists and classical optimists becomes unnecessary. Strawson offers separate arguments for each of these last two claims. Following Russell (1992), I will call these the "rationalistic argument" and the "naturalistic argument."

3. TWO ARGUMENTS FOR STRAWSONIAN MORAL RESPONSIBILTY

Strawson offers two separate arguments in favor of his theory of moral responsibility. These arguments separately show that a theory of moral responsibility that starts from the participant stance can fulfill two criteria for a theory of moral responsibility. First, the *rationalistic argument* shows that description and analysis of the reactive attitudes can allow one to reflectively distinguish morally responsible agents from those who are not morally responsible. This argument is not intended as a justification of Strawson's theory, but only as reason to believe that the theory can fulfill the condition that a theory of responsibility can allow us to distinguish morally responsible agents from those agents who are not morally responsible. I present my reconstruction of this argument in section 3.1 along with my argument that Strawson is a cognitivist expressivist about distinguishing responsible from non-responsible agents. Second, the *naturalistic argument* shows that external justification for holding responsible is unnecessary. Rather, Strawson claims that our practices involving holding each other morally responsible are justifiable from within those practices themselves. I present my reconstruction of this argument in section 3.2.

3.1 The Rationalistic Argument

First, the *rationalistic argument* proposes that the truth or falsity of determinism is irrelevant to an adequate theory of moral responsibility. To establish this point, Strawson offers a description and analysis of circumstances that are commonly thought to make reactive attitudes appropriate or render them inappropriate. First, it is necessary to note an important characteristic of the reactive attitudes. According to Strawson, these attitudes "involve, or express, a certain sort of demand for inter-personal regard. The fact of injury constitutes a prima-facie appearance

of this demand's being flouted or unfulfilled" (1962, p. 195). In general, assistance or injury constitutes an initial appearance of a beneficent or offensive will with respect to some interpersonal demand. When this happens, a reactive attitude, possibly gratitude or resentment, is stimulated in response to the good or ill will of the benefactor or offender. Sometimes, though, in our ordinary interpersonal dealings with one another, we discover some further fact that inhibits our initial reaction. At this point, in order to anticipate my argument in defense of expressivism, it is important to emphasize the proposal that reactive attitudes are psychologically prior to determination of the responsibility of an agent. Only after or concurrent with a reactive attitude does the question of the responsibility of an agent arise to potentially inhibit that attitude. Strawson distinguishes two categories of cases in which such inhibition seems appropriate.

Consider a situation in which someone, apparently at random, hits you in the face as you are walking down the street. You become resentful toward that person, and blame her, maybe even retaliate against her. Two different situations might inhibit this response. On the one hand, you might later realize that the injury was an accident, or maybe the person was pushed. In this case, it is likely that your initial reaction to that particular *action* becomes inhibited, and you might even apologize for your misplaced blame. On the other hand, you might later realize that the *person* was under exceptional stress or manipulation, or perhaps the person turns out to be child or a "hopeless schizophrenic" (1962, p. 188-189). In this case, a fact about the person, not just the particular action, is likely to inhibit your reaction, and render blame or punishment inappropriate. Following Watson (1987), I will call the first category of exculpatory conditions, those which relate to actions, cases of *excuse* and the second category, those which relate to persons, cases of *exemption*.

In the former class of cases, the person remains a participant in the moral community – one to whom interpersonal demands are addressable and for whom the reactive attitudes are appropriate. Excuse for some action is appropriate when there is evidence that the actor has no particularly beneficent or offensive will, when there is no unusual respect for or violation of some demand. In the latter class of cases, however, the agent does not remain a participant in the moral community. It is not consideration of the act itself that leads us to exculpate such a person, but consideration that the agent is not the sort of agent to whom interpersonal demands are addressable. Such a person is exempt from participation in the moral community either *at the time of injury*, in the case of stress or manipulation, or *in general*, in the case of a child or psychologically abnormal individual of some sort. Exempt persons may evince beneficence or offence; but we naturally and appropriately look on cases of exemption with an objective attitude in much the way pessimists and classical optimists inappropriately look on all human beings with an objective attitude. Punitive and approbative attitudes or practices are inappropriate for such agents because they are incapable of recognizing or responding to the demands placed upon them. For this reason, our demands toward them tend to be inhibited, and we are led to look on such agents as individuals to be trained or controlled.

Cases of exemption are particularly significant because pessimists effectively hold that if determinism is true, then all human beings are exempt. However, even without consideration of determinism, there are clear sets of paradigmatic exemption cases, in which it is judged appropriate to exculpate an agent for psychological or developmental features that indicate an abnormal ability to recognize and respond to interpersonal demands. This fact indicates that we have no reason in practice for allowing determinism to count as an added consideration. The truth or falsity of determinism does not, and need not, play a role in determining whether an

agent is morally responsible on this account of moral responsibility, so parsimony dictates that we leave it aside. Even if we accept the truth of determinism, Strawson points out, it is a self-contradiction to claim that abnormality is the universal condition (1962, p. 191). Therefore, on this account of moral responsibility, determinism is a superfluous consideration that is unnecessary for a theory of moral responsibility. From description and analysis of the reactive attitudes alone, we can distinguish morally responsible from non-morally responsible agents.

Still, there is an ambiguity in this last statement that must be addressed. On the one hand, an expressivist might hold that the reactive attitudes provide an adequate theory of responsibility because they determine when it is appropriate to hold an agent morally responsible, but do not imply either the truth or falsity of judgments about the responsibility of an agent. According to this interpretation, expressivism is *non-cognitivist* about moral responsibility judgments because it holds that, for all X, the claim "Agent X is morally responsible" is neither true nor false. On the other hand, an expressivist might hold that the reactive attitudes determine when it is appropriate to hold an agent morally responsible precisely because these attitudes track the truth or falsity of judgments about the responsibility of that agent. According to this latter interpretation, expressivism is *cognitivist* about moral responsibility because it holds that, for all X, the claim "Agent X is morally responsible" is true or false. It follows from the non-cognitivist interpretation of expressivism that any given agent can both count as morally responsible and fail to count as morally responsible in different contexts without contradiction. Because expressivists base determinations of responsibility in the reactive attitudes, and propositional formulations of these judgments may be contradictory, it is impossible for a non-cognitivist expressivist to reflectively determine whether some agent is morally responsible or not.

Some interpreters, for example Vargas (2005), refer to Strawson as a "clear case" of non-cognitivism about moral responsibility. When we consider that, for Strawson, negative and positive reactive attitudes respectively constitute blame and praise, it seems as though the non-cognitivist interpretation is correct. Reactive attitudes happen quickly and automatically, and no belief or other form of propositional content is obviously involved. Without propositional content, it is not clear how a reactive attitude could be construed as true or false, so it is not obvious how a contradiction among reactive attitudes could arise. However, the story Strawson presents to explain how the reactive attitudes can provide an adequate theory of responsibility does not end with the reactive attitutdes. He goes on to describe the conditions under which we inhibit these attitudes, and judge it appropriate to do so. This part of the story, which describes conditions for exculpation, provides the theory that allows us to reflectively determine when an agent is responsible and when not. For this reason, exculpation conditions are essential for an adequate expressivist theory of responsibility. Furthermore, the varieties of exculpation, both excuse and exemption, entail a judgment about an action or an agent that does have propositional content. For example, "Agent X was pushed" or "Agent X is a child" are both reflections that are either true or false. Therefore, while Strawson may be a non-cognitivist about the initial reactive attitudes, he is a cognitivist about moral responsibility. This point will become particularly important as we consider the objections to Strawsonian expressivism in section 4.

The rationalistic argument proposes that description and analysis of the reactive attitudes toward some agent are sufficient for determining whether or not that agent is morally responsible. There is no reason to believe that this view is a non-cognitivist one, and good reason to believe that we should interpret it as a cognitivist position about moral responsibility judgments. In the next section, I turn to the reasons Strawson offers for believing that this view is

justified, even without the external justification sought by optimists and pessimists about responsibility.

3.2 The Naturalistic Argument

The second argument Strawson presents in favor of his theory, the *naturalistic argument*, proposes that inclusion of the human attitudes in a theory of responsibility renders the kind of justification for our practices sought by pessimists and classical optimists unnecessary. Although the rationalistic argument shows that consideration of determinism is *not required* for an adequate theory of responsibility, a pessimist or classical optimist might respond that this argument does not show our practices or attitudes are *justified*. The rationalistic argument does not by itself defend against the threat determinism poses to our punitive and approbative practices and attitudes. However, pessimists and classical optimists alike assume that justification of our punitive and approbative practices requires that we have some reason *outside of our everyday practices* for ignoring determinism (for example, because there is reason to believe that determinism is false or because our practices are useful for regulating behavior even if determinism is true). Thus, on their view, justification of our practices requires us to explain why determinism does not exempt all human beings, thereby forcing us always to take the objective stance toward everyone.

In response, Strawson points out that even if the truth of determinism *should*, from the objective standpoint, exempt all human beings from moral responsibility, it is impossible in practice for us to universally exempt each other in this way. Participation in normal interpersonal relationships is part of our nature, and "[such participation] precisely is being exposed to the range of reactive attitudes and feelings that is in question" (1962, p. 192). In turn, being an

appropriate object of the reactive attitudes implies moral responsibility. Therefore, we are committed, not just to ordinary interpersonal relationships, but also to the human attitudes that constitute holding each other responsible. It is impossible, in practice, to universally give up these attitudes even if a theoretical consideration such as determinism suggests that we should. Of course, we can *sometimes* abstain from the participant stance, withdrawing from the human attitudes, in favor of an objective standpoint. The objective stance is useful, even with respect to normal adults who are capable of recognizing interpersonal demands, "as a refuge, say, from the strains of involvement; or as an aid to policy; or simply out of intellectual curiosity" (1962, p. 190). Still, Strawson claims, it is not possible to completely withdraw from the human attitudes. The expression of human nature at least partially requires exposing oneself to the human attitudes in the participant stance. Universally taking the objective stance would mean giving up our very humanity. In a footnote, Strawson draws a useful comparison with Hume's justification of induction:

> Compare the question of the justification of induction. The human commitment to inductive belief-formation is original, natural, non-rational (not irrational), in no way something we choose or could give up. Yet rational criticism and reflection can refine standards and their application, supply 'rules for judging of cause and effect'. Ever since the facts were made clear by Hume, people have been resisting acceptance of them (1962, p. 204).

The analogy between reactive attitudes and perceptual experience is apt given that both are automatic and natural, yet open to misplaced dismissal.

The fact that our human nature commits us to the reactive attitudes indicates that the pessimists and classical optimists are wrong to seek external justification for our practices, or the attitudes on which they are founded, by considering the human condition *only or originally from an objective standpoint*. These philosophers, like politicians or bureaucrats, take the objective stance "as an aid to policy." However, rather than starting with an understanding of interpersonal

human involvement and using the objective stance only as an *aid*, pessimists and classical optimists go wrong by starting with concern for the effect determinism should have on punitive and approbative practices and forgetting the human attitudes entirely. To these disputants, determinism seems pressing since they have only a partial view of the facts in mind – namely, our punitive and approbative practices, and not the attitudes on which these practices are founded. Yet, rather than turn to an internal analysis of the human attitudes, pessimists and classical optimists "over-intellectualize," seeking external justification for our practices and attitudes from the objective standpoint.

The reactive attitudes can alleviate concern with determinism both because they are adequate for distinguishing responsible people from those who are not and because they are inescapable for human beings *qua* human beings. The rationalistic argument shows that we can distinguish morally responsible agents from cases of excuse and exemption by describing and analyzing when and how reactive attitudes arise in response to actions and agents within the participant stance. The naturalistic argument shows that, regardless of what we conclude in the objective stance, we are not able to give up all instances of our reactive attitudes in order to view others and ourselves from a purely disinterested standpoint. Therefore, the demand for external justification for our punitive and approbative practices and attitudes asks too much of a theory of moral responsibility. The point of these arguments is to show that reconciliation in the debate surrounding determinism can and should come by reminding the interlocutors about the reactive attitudes, and further suggesting that they take the participant stance as a starting point for constructing a theory of responsibility. Since consideration of determinism, and objective standpoint considerations generally, are both superfluous for a theory of responsibility (according to the rationalistic argument) and inefficacious in undermining the fact that we hold

each other responsible (according to the naturalistic argument), we should leave them aside. At this point, there is no further basis for the disagreement between pessimists and classical optimists. Still, there are important criticisms of Strawson's view in the contemporary literature. Given my reconstruction of Strawson's two arguments for, and my cognitivist interpretation of, his position, I hope to dispel two of these contemporary criticisms in the remainder of this essay.

4. LUCK AND REVISION

While Strawsonian expressivism is widely influential among contemporary theorists of moral responsibility, many important objections have been raised. Here, I consider and respond to two representative examples: one metaphysical objection and one epistemological objection. First, a more concise presentation of Strawson's *main argument for expressivism*:

1. Description and internal analysis of the human attitudes can provide an epistemologically adequate theory of responsibility (Rationalistic Premise).

2. Natural commitment to the human attitudes renders external justification of these attitudes unnecessary (Naturalistic Premise).

   Therefore,

3. Moral responsibility may be understood and justified using description and internal analysis of the human attitudes alone.

1 and 2 are, respectively, the conclusions of the rationalistic and naturalistic arguments. They are also the premises of the main argument for expressivism, so I will refer to them as premises from this point on. Note that premise 1 and 2 depend on each other. Without premise 1, even natural commitment to punitive and approbative practices and attitudes might not allow us to determine who is responsible and who is not. Without premise 2, the practices and attitudes that allow us to distinguish responsible agents from agents who qualify for excuse and exemption might not be justified. The epistemological objection argues that premise 1 is false because consideration of the possibility of moral luck shows that we cannot use description and analysis of the reactive attitudes to adequately distinguish morally responsible from exempt agents. The metaphysical objection argues that premises 1 and 2 are in conflict with one another, and accepting the revisability of our initial reactive attitudes shows that we should give up premise 2.

4.1 Moral Luck

The first objection is raised against the rationalistic premise by Watson (1987). Strawson intends this first premise to imply that the truth or falsity of determinism is irrelevant to an adequate theory of responsibility. In fact, I will construe this premise more broadly to imply that no objective standpoint criteria – those which are fixed independent of description and analysis of the reactive attitudes – are required to formulate a theory of responsibility that allows us to reflectively determine which agents are responsible and which are not.[4] A Strawsonian must be able to show that we can have an adequate theory of responsibility without appeal to some external criteria, such as agent causation or consequentialism. To avoid being external in the relevant sense, the criteria that allow us to reflectively distinguish responsible from non-responsible agents must be formulated as hypothesis based on the evidence provided by the reactive attitudes for the purpose of testing them against reactive attitudes which arise in further cases. To point us in the right direction, Strawson offers some paradigmatic cases of excuse and exemption, and a description of how a child or psychoanalytic patient emerges from exemption into full responsibility. He also suggests that, in the case of excuse, empirical evidence pertaining to the intention of an agent, or, in the case of exemption, the cognitive capacities of an agent, are what causes inhibition of reactive attitudes toward that agent. Although there are paradigmatic cases which fit these criteria and seem to track our reactive attitudes, expressivism remains open to incorporating other criteria if counter-examples show our reactive attitudes do not always track intention or cognitive capacity. However, if it turns out that reactive attitudes sometimes conflict with respect to some agent in a way that is not amenable to a coherent set of criteria, then as a cognitivist I accept that expressivism fails to formulate an epistemologically adequate

[4] Other examples of objective standpoint criteria might include the falsity of fatalism or of God's foreknowledge.

theory of responsibility based on reactive attitudes in the participant stance. If expressivism does fail to allow us to distinguish morally responsible from non-morally responsible agents, premise 1 of the main argument is unacceptable.

Watson (1987) cites a case that he believes undermines formulation of a coherent expressivist interpretation of Strawson's theory by prompting conflicting reactive attitudes toward the same agent, therefore leaving it indeterminate whether or not that agent is morally responsible according to expressivism. To begin, he notes, "reactive attitudes are sensitive not only to the quality of others' wills, but depend on a background of beliefs about the objects of those attitudes" (1987, p. 121). In cases of excuse, a belief about how an action came about defeats any reason to believe the will of the actor runs contrary to the demands which that actor is expected to respect. In contrast, exemption takes place when reactive attitudes are inhibited by beliefs that imply an agent does not have the capacity to be addressed as a member of the moral community. The problem is that those who are incapacitated in this way can sometimes show good or ill will towards others, thereby stimulating reactive attitudes. Since, according to expressivism, reactive attitudes are supposed to determine when an agent is or is not responsible, it is difficult to see how we can reflectively recognize when an agent is exempt. Pessimists or classical optimists can avoid this problem by arguing that their theoretical criteria, external to the reactive attitudes, determine when an agent is properly considered exempt or responsible. This response is not available to the Strawsonian because appropriateness for the reactive attitudes constitutes moral responsibility, and appropriateness can only be determined by analyzing when the reactive attitudes are stimulated or inhibited. If the beliefs that inhibit reactive attitudes in the paradigmatic cases of exemption imply objective standpoint criteria for moral responsibility, then exemption cases undermine expressivism. Watson initially avoids this problem for two

cases of exemption, children and agents "under great strain," by showing that exemption in these cases depends only on how the reactive attitudes are modified. He then argues that one case of exemption, "being unfortunate in formative circumstances," is inimical to expressivism because it is impossible to know who fits this description without appeal to independent criteria for moral responsibility.

First, consider how Watson shows that exempting children is amenable to expressivism. Children are members of Strawson's second sub-type of exemption cases, which exempt agents in general rather than just at the time of injury. The fact that children lack the cognitive capacity and moral competence required to understand and respond to the demands placed upon them explains their exempt status. Furthermore, children only gradually become members of the moral community because they only gradually develop the required capacities and competence – possibly recognizing different demands at different points in time on an individual basis. Since good will with respect to some demand requires recognition of and ability to respond to that demand, expressivism can accommodate the fact that reactive attitudes become inhibited in response to a belief that children often lack such recognition and only gradually develop it.

Second, consider how Watson shows that exempting agents due to their "being under great strain" is amenable to expressivism. Such agents fall under Strawson's first sub-type of exemption cases, which exempt agents only at the time of injury. In these cases, the agent in question is generally able to recognize and respond to demands; so, unlike the case of children, these cases cannot be explained by lack of these capacities. Rather, the fact that agents under great strain often act uncharacteristically explains their exemption. Demand for good will is limited to *normal* circumstances, and provides for exception under unusual circumstances. For apparent good will with respect to some demand to count as beneficence, reflective endorsement

of that will is required. Expressivism can accommodate the fact that reactive attitudes become inhibited in response to a belief that someone is not expressing their "true self" because this condition coheres with the general requirement that responsible agents are capable of recognizing and responding to interpersonal demands.

For expressivism, no contradiction arises among the principles required to explain the exemption of children and people under great stress; so it is not necessary to appeal to agent causation, social regulation, or any other external theoretical requirement. Again, the principles hypothetically formulated to exclude children and persons under great stress from responsibility are not formed independent of the reactive attitudes, and are therefore internal to the attitudes and practices involved in holding responsible. These principles are intended to track the reactive attitudes and they may be tested against intuitions about other cases. However, Watson returns to another exemption case of the second sub-class, which he does think requires external criteria in order to coherently exclude from moral responsibility: those who were "unfortunate in formative circumstances." Watson quotes at length a description of the life of Robert Harris as a case study of this type of exemption. In brief, Harris egregiously murdered two boys in the course of a 1978 bank robbery, laughed about the crime, and expressed interest in killing police who would later arrive at the scene. Indignation toward Harris was magnified by his casual disdain for even basic social norms and contempt for human life. By all accounts, Watson claims, Harris was the epitome of evil, "an 'archetypal candidate' for blame" (1987, p. 128).

After arriving at this conclusion, Watson considers how our reactive attitudes are affected by the developments that lead someone like Harris to become evil. It turns out that Harris had a particularly brutal childhood. His mother admits she feels guilty that she was never able to love him; she blamed Robert for her problems with her husband; he would plead for affection and get

tossed aside; both his parents physically abused him; and as a teenager he was raped several times. His sister is quoted in the *Los Angeles Times* article that Watson excerpts for his description:

> [Barbara Harris] put her head in her hands and cried softly. 'One killer out of nine kids…The sad thing is he was the most sensitive of all of us. When he was 10 and we all saw 'Bambi', he cried and cried when Bambi's mother was shot. Everything was pretty to him as a child; he loved animals. But all that changed; it all changed so much." (1987, p. 129)

It seems our indignation for Robert Harris the man is modified by the belief that Robert Harris the boy was "unfortunate in formative circumstances."

In the case of a child, inhibition of reactive attitudes is explained by knowledge that children have underdeveloped cognitive capacities and lack of moral competence. In the case of people experiencing abnormal stress, evidence that they are incapable of normally responding to interpersonal demands explains why reactive attitudes are inhibited. Initially, one might think there is some fact about the way Harris was as an adult that exempts him, just as there are non-historical exempting facts about children and people under stress, so that it is not necessary to appeal to the historical conditions that made Harris into the kind of man he became. For example, Harris apparently does not respond to moral demands at all, and his outright rejection of the moral community might mean moral demands are not addressable to him in the way required for moral responsibility. The problem with this explanation is that our reactive attitudes are not inhibited by the fact that he is not morally addressable at the time of his crimes, nor do they respond to any other structural or situational facts about Harris. If we exempt Harris for his outright rejection of the moral community, then modification of reactive attitudes cannot be the ground for this exemption; the ground must instead be some independent theoretical conviction. Of course, this approach is not amenable to expressivism. Moreover, if we exempt Harris for his

radical rejection of moral demand, then all radically evil agents must also be exempt from responsibility simply because of the extremity of their rejection. Without a sympathetic historical background, this fails to conform to our normal reactive attitudes. The more obvious consideration, which does inhibit our attitudes toward Harris, is the abusive history leading up to rejection of the moral community.

Although knowledge of an abusive childhood modifies our reactive attitudes toward an evil person, according to Watson this consideration also implies an independent theoretical conviction, which is unacceptable for expressivism. First, note that Harris does not meet the expressivist conditions for excuse: his history does not undermine the thought that he acted maliciously. Second, although it is not obvious whether Harris meets exemption conditions, if we accept that he was incapacitated for membership in the moral community by his unfortunate childhood, then, once again, Watson argues there is reason to exempt all radically evil agents simply for being radically evil (1987, p. 133). To see why, consider a radically evil agent for whom there is no apparent historical explanation, someone who is inexplicably vicious. Watson calls such individuals "bad apples." Even when there is no obvious abuse or other social explanation for the viciousness of a bad apple, Watson suggests that we should suppose there is something that makes bad apples incapable of recognizing and responding to moral demands, for example, something "in their genes or brain." Supposing there is such an explanation for every radically evil individual, this explanation should play the same role in exempting the individual that abuse plays in exempting Harris. With respect to Harris, our reactive attitudes fluctuate between antipathy and sympathy. Therefore, our reactive attitudes toward all radically evil agents also should be ambivalent, leaving it indeterminate whether such agents are morally responsible or not on the expressivist theory.

This argument presents one part of the epistemological problem that historical considerations pose for expressivism. Since ambivalent reactive attitudes make it indeterminate whether radically evil agents are morally responsible or not, the reactive attitudes are not amenable to the formulation of a coherent set of principles for deciding whether radically evil agents are responsible. However, Watson extends this problem again from radically evil agents to all agents, evil or not. Just as we are largely ignorant of the historical considerations that likely make radically evil agents the way they are, we are largely ignorant of the historical considerations that make anyone, including ourselves, the way we are. When we turn the sympathy we feel for Harris as we consider his history toward other people, or inward toward ourselves, we recognize that the difference between "us good people" and "those evil people" is just a matter of moral luck. In this way, reactive attitudes toward all people are made ambivalent by ignorance of the historical factors that determine what kind of person one becomes. This ignorance, Watson supposes, should make us skeptical of the reactive attitudes generally (1987, p. 137). Furthermore, for an expressivist, general skepticism of the reactive attitudes is not separable from skepticism about responsibility. The point of formulating a reflective theory of responsibility is that such a theory will allow us to correct our reactive attitudes where an error arises here or there; but if all our reactive attitudes are suspect, then there is no expressivist basis on which to formulate such a theory. If Watson is correct to generalize skepticism about reactive attitudes, then we must give up premise 1 of the main argument.

In response to Watson, I raise three objections. First, the Harris case does not exemplify a pernicious ambivalence in our reactive attitudes. Although we are variously antipathetic and sympathetic toward Harris, this opposition in our attitudes is not the one Strawson claims constitutes the opposition between blame and praise. It would be problematic for expressivism if

we experienced both indignation and approbation with respect to Harris, or some similar sentiments, since these are the relevant contradictory reactive attitudes. It would also be problematic for expressivism if our reactive attitudes toward Harris indicated that we both blame him and exempt him. Our sympathy with Harris, however, does not contradict our indignation. According to the cognitivist interpretation of Strawsonian expressivism I offered, a belief that contradicts the belief that an agent is able to understand and respond to demands would inhibit reactive attitudes, but mixed feeling about an agent does not result in exemption. Since there is no contradiction in "sympathetic blame," expressivism of the cognitivist sort that I attributed to Strawson has no problem accounting for the ambivalence we feel as we learn about the past that made Harris malicious. While McKenna (1998, p. 206) accepts that Harris is not a member of the moral community in the course of his response to Watson, I see no reason to concede this point. Consider how you would respond to Harris if you met him in a dark alleyway. If you would not immediately run away, I suspect you would at least be on your guard in his presence. This does not change even when you know of his childhood. Such ostracization shows that negative reactive attitudes are not dispelled. It is important to remember that expressivism is a theory of social moral responsibility, not of legal responsibility. I noted earlier (see n.2) that even for an expressivist, legal punitive and approbative practices may require justification that goes beyond that required for our everyday attitudes and practices. It may be true that knowledge of an abusive childhood would change how the legal system should deal with a criminal – maybe we should treat rather than punish such a person – but such knowledge does not change our interpersonal attitudes and practices in the participant stance. Therefore, there is no reason to suppose that ambivalence with respect to Harris makes it impossible to formulate an adequate expressivist theory of responsibility.

Second, at this point the analogy between Harris and bad apples also breaks down. Since Harris is not exempt because of his abusive history, neither are bad apples exempt for some hypothetical malfunction in their genes or brain. It is instructive to note that, even if Harris were exempt, there is an important dissimilarly between cases like Harris and bad apples. Namely, Harris is a victim of his abusers, but it is impossible to be a victim of one's genes or brain in a sense which calls for reactive attitudes. Victimhood in this sense implies that someone has violated a demand that the victim legitimately places on others. The reason we sympathize with Harris to some degree is that the interpersonal violations he was subject to call for some punitive attitudes and practices. However, neither genes, nor brains, nor any other non-person is addressable with interpersonal demands, so it is wrong to characterize bad apples as victims. Of course, bad apples may be incapacitated in some way, for example, by a brain tumor; but then there is an explanation for the apparent viciousness, so such incapacity is epistemologically different from someone who is under hypnosis or high stress, or a psychologically deranged individual. In all of these cases it might take some investigation before evidence presents itself which leads you to inhibit your reactive attitudes, but there is no reason to inhibit reactive attitudes independent of such evidence. Even if one does not accept my argument that Harris is not exempt, the analogy between Harris and bad apples is flawed because bad apples are not victims like Harris. Therefore, there is no reason to think inexplicably vicious individuals are exempt, and they do not pose a problem for expressivism.

Third, the analogy Watson draws between bad apples and all people is also flawed. Even if you still think that Harris is exempt and the analogy with bad apples holds, these propositions do not imply that we should generalize skepticism about responsibility for all people based on historical considerations. Watson points to the problem of moral luck in order to impress the idea

that skepticism of negative reactive attitudes with respect to bad apples should generalize in this way (1987, p. 137). None of us creates our character out of nothing, so, Watson claims, it is a matter of luck whether we turn out virtuous or vicious. Since, for most people, and even for ourselves, we lack knowledge of the past that determines character, Watson argues that we, in general, should be skeptical of, and therefore inhibit, the reactive attitudes. However, lack of evidence for self-creation is not evidence of incapacity. Just like in the case of bad apples, there is no reason to exempt anyone without appropriate exculpatory evidence available. There is also a further dissimilarity between inexplicably evil persons and other members of the moral community that did not exist between Harris and bad apples. Normal people are different from bad apples because normal people show good or ill will variously in the course of their lives or even in the course of a day. Evidence that distinguishes good will from ill will, and persons who are capable of displaying these intentional states from persons who are not so capable, is all that is necessary to formulate a theory that distinguishes morally responsible individuals from excused or exempted ones. Even if it is true that luck plays a large role in how we become who we are, an adequate expressivist theory of responsibility is still possible. Justification for basing such a theory on the reactive attitudes is another matter, which I consider in section 4.2.

Although Strawson himself responds to the threat of determinism by noting that it is irrelevant to the formulation of an adequate theory of responsibility, it is possible for some other consideration to threaten general skepticism about the reactive attitudes. General skepticism about the reactive attitudes would undermine the ability to formulate a theory of responsibility based on those attitudes, and would have to lead to skepticism about moral responsibility. Watson attempts to engender this skepticism by showing that historical considerations, particularly the idea that luck plays a large role in who we become, should lead to skepticism

about the reactive attitudes. In response, I argue that (a) understanding the history of Harris does not exempt him; (b) even if his history did exempt him, failure to understand the evil of bad apples does not show that we should exempt them; and (c) even if we should exempt bad apples, the fact that normal people are not pervasively evil like bad apples is a relevant difference, which shows that we need not accept general skepticism of the reactive attitudes. Thus, Watson does not show that we should give up premise 1 of the main argument for expressivism.

Strawson's claim that the reactive attitudes are an adequate basis for distinguishing responsible and non-responsible agents is left untouched by the Harris case, bad apples, or moral luck. However, if we interpreted Strawson as a non-cognitivist about moral responsibility judgments, then the tension between our antipathetic and sympathetic response to Harris at various points would be enough to undermine our ability to reflectively determine whether or not Harris is responsible. In contrast, according the cognitivist interpretation of Strawson, sympathy (or any other positive emotion) toward some agent is not enough to exculpate. Rather, exculpation is specifically an inhibition of the original reactive attitude by a belief about the exculpated individual that contradicts the presumed capacity for understanding and responding to moral demands. The Harris cased does not exemplify any such contradiction, and neither do bad apples nor lucky persons display any such incapacity. Still, one point Watson raises remains unopposed. At the end of his article, Watson claims that premise 2 – the claim that we need not appeal to external justification – is also false. He notes that Albert Einstein might be a good example of someone who gave up the reactive attitudes all together, so it might not be true that we are committed to these attitudes, as Strawson claims. The argument against premise 2 is more fully fleshed out by Paul Russell, and it is this argument that I will turn to next.

4.2 Reactive Attitude Revision

The second objection to expressivism is raised by Russell (1992), who argues that we should give up premise 2 of the main argument for expressivism. To begin, Russell claims that Strawson either misses the point of the pessimistic position or a conflict arises between the rationalistic and naturalistic premises, which renders the argument incoherent. In either case, it turns out there is still reason to doubt the metaphysical claim that some morally responsible agents exist and, consequently, to ask for justification for our practices and attitudes from the objective stance. According to this argument, Strawson equivocates on the definition of naturalism with respect to the reactive attitudes.[5] On the one hand, naturalism about reactive attitudes might mean that persons are committed to these attitudes at the type level; on the other hand, naturalism might mean that persons are committed to these attitudes at the token level.

Type-level naturalism holds that human beings are necessarily disposed to experience the reactive attitudes. Russell acknowledges that, interpreted as a type naturalist, Strawson is correct to point out that we are committed to the reactive attitudes in interpersonal relations. However, pessimists need not be skeptical about dispositional commitment to such attitudes. Rather, pessimists only need to claim that for any token reactive attitude that arises, it is possible to inhibit that attitude. This latter position contrasts with token-level naturalism, which holds that persons are committed to at least some instances of the reactive attitudes that in fact arise. Unlike token-level naturalism, type-level naturalism does not imply commitment to any particular reactive attitude that might arise, so type-level naturalism is compatible with pessimism. Since Strawson intends for his argument to undermine pessimism, it is more charitable to read him as a

[5] For more on Strawson's conception of naturalism see Strawson (1985).

token-level naturalist. However, Russell argues, token-level naturalism is inconsistent with the rationalistic argument.

Recall that the rationalistic argument depends on recognition of our ability to excuse or exempt some agents in light of further facts and subsequently revise reactive attitudes toward them. Since Strawson argues that any particular instance of the reactive attitudes can be inhibited by further facts, it seems problematic for him also to hold that we are naturally committed to any token reactive attitudes that arise. Remember, both Strawson and the pessimists agree that incapacity of some sort makes an agent inappropriate for the punitive and approbative practices and attitudes. The pessimist further claims that determinism generates one such incapacity for every agent, so all particular instances of reactive attitudes should be inhibited.[6] Thus, Strawson must either miss the point of the pessimistic worry (as a type naturalist) or contradict the rationalistic argument by holding that at least some instances of the reactive attitudes are unrevisable (as a token naturalist). In either case, the result is that pessimists and classical optimists are right to think our punitive and approbative practices are not *internally* justifiable, instead requiring justification from the objective standpoint. Thus, it seems that premise 2 of the main argument is undermined by an implication of premise 1, that reactive attitudes are revisable in light of further facts.

While I think Strawson commits himself to both kinds of naturalism, this objection fails because, according to the cognitivist interpretation of expressivism, token naturalism is not actually inconsistent with the rationalistic argument. Strawson is committed to both of the following claims:

[6] Some might claim that there are other considerations that generate universal incapacity, e.g. no one has the capacity to make themselves who they are (G. Strawson, 1994). I believe my response in the following paragraph incorporates all such considerations.

1. *All* token reactive attitudes have the *potential* to be revised in the face of further facts (from premise 1 of the main argument).

2. *Some* token reactive attitudes are *actually* inescapable (from premise 2 of the main argument).

These claims are not contradictory because the reactive attitudes that are in fact modified by further facts are not the same reactive attitudes as the ones that are inescapable. The further facts which actually (and ought to) *excuse* an action are facts that undermine the belief that the action was done with a beneficent or malicious intent. The further facts which actually (and ought to) make an agent *exempt* are facts that make that agent incapable of recognizing or acting on interpersonal demands. Again it is important to remember that, according to cognitivist expressivism, the truth-maker for the claim that some action is excusable or that some agent is exemptible is a fact about the action or the agent, not simply a change in valance of the reactive response toward the agent. Although any reactive attitude is potentially modifiable upon recognition of these exculpatory conditions, not all actions or agents fit the conditions for excuse or exemption. There is no contradiction between claims 1 and 2 unless we antecedently assume that all token reactive attitudes not only have the potential for revision, but also should in fact be revised.

Furthermore, premise 1 of the main argument alone implies that there is no reason to commit to this antecedent assumption. According to premise 1, which Russell provisionally accepts, there are two possible reasons for universal exculpation: universal excuse for actions and universal exemption for persons. It is obvious that some actions are done with intent to help or harm and some are not, so there cannot be universal excuse. It is less obvious whether or not there is reason to universally exempt. To emphasize that universal exemption is possible, Russell stresses the point that Strawson waivers between calling exempted agents “incapacitated” and

calling them “abnormal” (1992, p. 153). Strawson rightly notes that any thesis which holds ‘abnormality is a universal condition’ is self-contradictory (1992, p. 191). Yet, Russell points out, it is not self-contradictory for incapacity to be the universal condition. What Russell fails to recognize is that, in the participant stance, incapacity just means failure to function normally with respect to recognition and response to interpersonal demands. Any incapacity that applies universally would make it impossible to distinguish in practice those agents who are responsible and those who are not. Therefore, by accepting premise 1 (as I have argued we should), Russell blocks his own criticism because premise 1 implies that universal exemption is impossible. Since, if true, determinism applies to everyone equally, it cannot have any bearing on what constitutes incapacity in the sense operative for Strawson. Participation in interpersonal relationships necessarily involves placing demands on each other, so according to expressivism some morally responsible agents must exist. The fact that token instances of the reactive attitudes are modifiable by rational considerations does not, by itself, imply that we should inhibit all instances of these attitudes.

Again, Russell argues that the rationalistic and naturalistic premises conflict and, since the reactive attitudes are revisable, as the rationalistic premise suggests, a Strawsonian expressivist should give up the naturalistic premise, asserting that justification requires rejection of determinism and other potential universal exemptions from the objective standpoint. In response, I argue that Russell begs the question. The reason he offers to believe we are not committed to any token reactive attitudes is that all token reactive attitudes are revisable; but the only reason to believe all token reactive attitudes are revisable is that we are not committed to any of them. Moreover, I argue, we are committed to precisely the reactive attitudes for which there is no contravening evidence. Since the operative notion of incapacity in the participant

stance does not allow for universal exemption, it is not possible to give up all token instances of the reactive attitudes. Taken together, these points refute the objection Russell offers.

5. CONCLUSION

The main argument Strawson gives for expressivism rests on two claims. First, he claims we can know who is responsible and who is not by description and analysis of the reactive attitudes. Second, he claims that participation in social interaction commits us to reactive attitudes that arise in response to the good or ill will of those who can recognize and respond to the demands of the moral community. After reconstruction the two arguments Strawson offers for these claims, I have attempted to defend each against two common objections.

The moral luck objection argues that it is impossible to formulate a theory based on the reactive attitudes that epistemologically allows us to determine which agents are the morally responsible ones. Sometimes people argue that objective standpoint considerations such as determinism, fatalism, or God's foreknowledge might undermine moral responsibility. Strawson attempts to avoid such considerations by showing that we can have an adequate theory of responsibility by analyzing and forming principles based on the reactive attitudes which arise in the participant stance. The criticism offered by Watson is interesting because he attempts to show that we cannot formulate a theory by analysis of the reactive attitudes alone. I argue that Watson fails to show this both because ambivalent sentiments toward Harris are not inconsistent with cognitivist expressivism and because the analogies that Watson proposes to lead us toward radical skepticism about moral responsibility in the participant stance fail.

The revisionist objection argues that we are not justified in formulating a theory of responsibility based in the reactive attitudes alone because we are not committed to the reactive attitudes in the way suggested by Strawson. Russell argues that the rationalistic premise shows that we are not committed to token reactive attitudes, as implied by the naturalistic premise, and we are therefore not committed to the existence of some morally responsible agents. This

argument implies that justification for belief in morally responsible agents must come from principles based not in the reactive attitudes, but in principles formulated from the objective stance, independent of the reactive attitudes. In response, I argue that we are in fact committed to some token reactive attitudes, namely, those for which we have no exculpatory evidence. According to the cognitivist interpretation of expressivism, exculpation requires a belief that an agent is incapacitated for moral address. Therefore, absent such a belief, no revision is required. The fact that there are cases in which no such belief arises is, as Strawson suggests, all the justification necessary for the claim that our punitive and approbative attitudes and practices establish the existence of some morally responsible agents.

The rationalistic premise establishes that it is possible to formulate a theory of moral responsibility based on the reactive attitudes. The naturalistic premise establishes that we are justified in formulating such a theory without rejecting all of the possible factors that seem to universally exculpate agents from the objective stance. Thus, the premises of the main argument for expressivism depend on each other by establishing two points necessary for an expressivist theory of moral responsibility: (a) it is possible to formulate principles based on the reactive attitudes and, (b) the reactive attitudes are a justified basis for the principles. The process of formulating an expressivist theory of responsibility may thus proceed by hypothesizing principles of moral responsibility based in the reactive attitudes that are evident in some cases, and testing these principles against the reactive attitudes which arise in other cases.

Strawson and Watson, along with many others who sympathize with expressivism, use this method, which is often referred to as *reflective equilibrium*, by appealing to their own intuitions and anecdotal examples of reactive attitudes observed in others. Yet, it should be clear that the input for an expressivist reflective equilibrium is not just the reactive attitudes of a few

academics, but the reactive attitudes of the folk in general. The most important implication of expressivism is that it suggests a conservative approach to the everyday practices and attitudes of normal people in everyday interpersonal situations. This may sound troubling at first, since common sense intuitions and practices may seem inconsistent and insensitive. However, this pretense should be resisted prior to empirical investigation of folk practices and reactive attitudes, especially if Strawson is right about the errors generated by objective stance theorizing about moral responsibility. Even if common practices and attitudes are not perfectly consistent, expressivism allows for some revision along Strawsonian lines (Vargas, 2004). Before taking this step, however, it is important to have an accurate empirical description of common practices and attitudes in hand for analysis. Fortunately, there is a growing body of evidence about folk intuitions and judgments about responsibility to draw from. Further research should take this evidence into account, including both the possibility that ambivalent reactive attitudes might challenge the expressivist theory and the possibility that inconsistent reactive attitudes might require revision.

REFERENCES

McKenna, M. & Russell, P. eds. (2008). *Free Will and Reactive Attitudes: Perspectives on P.F. Strawson's "Freedom and Resentment"* Aldershot, UK: Ashgate Press.

McKenna, M. (1998). The Limits of Evil and the Role of Moral Address: A Defense of Strawsonian Compatibilism. *Journal of Ethics*, 2, 123-142.

Nagel, T. (1979). "Moral luck." In T. Nagel (Ed.), *Mortal Questions* (pp. 24-38). Cambridge, UK: Cambridge University Press.

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

Strawson, P. F. (1962) Freedom and resentment. Proceedings of the British Academy, 48, 187-211.

Strawson, P. F. (1985). Skepticism and Naturalism: Some Varieties. New York: Columbia University Press.

Strawson, G. (1994). The Impossibility of Moral Responsibility. *Philosophical Quarterly*, 75(1-2), 5-24.

Vargas, M. (2004). Responsibility and the Aims of Theory: Strawson and Revisionism. *Pacific Philosophical Quarterly*, 85(2), 218-241.

Vargas, M. (2005). The Revisionist's Guide to Responsibility. *Philosophical Studies* 125(3), 399-429.

Watson, G. (1987). "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In F.D. Shoeman (Ed.), *Responsibility, Character, and the Emotions* (pp. 256-286). Cambridge, UK: Cambridge University Press.