# SITUATIONISM AND MORAL RESPONSIBILITY: AN EXTERNALIST ACCOUNT

MICHELLE CIURRIA

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PHILOSOPHY
YORK UNIVERSITY
TORONTO, ONTARIO

AUGUST 2014

**Abstract**

Situationism is the position that there is no such thing as broad, situation-invariant character. This view emerged from situationist psychology experiments, such as the famous Milgram Obedience Experiments. It is supposed to impugn character-based theories that posit motivationally self-sufficient virtues, such as (one natural reading of) Aristotelian virtue ethics. While situationists agree on this point, a survey of the literature reveals that they defend inconsistent accounts of moral responsibility, ranging from radically revisionary to staunchly conservative. In my dissertation, I begin by focusing on three of the most preeminent situationists—John Doris, Gilbert Harman, and Philip Zimbardo—and show that their accounts of moral responsibility are inchoate, mutually incompatible, and sometimes incongruous with the central logic of situationist psychology. I also argue that the epistemological problem, i.e., the problem that we have very weak and unreliable access to our mental states, undermines traditional psychological accounts of moral responsibility, which depend on strong introspective access. This position is corroborated by cognitive science research into cognitive distortions such as confabulation, rationalization, and cognitive bias. In the constructive part of my dissertation, I propose a new, externalistic account of moral responsibility, based on Holmes' standard of the reasonable person (SRP), which measures responsibility by what a reasonable person would do in the defendant's circumstances. This account was developed as an alternative to traditional choice and character-based theories of legal responsibility, and is particularly apt for assessing cases of negligence and coercion, which traditional theories were incapable of explaining. If we see situational constraints as a form of coercion, this account is also adept at capturing these types of excuses. I argue that an equivalent to SRP is long overdue in moral philosophy, for the purposes of capturing non-psychological excusing factors and accommodating empirical research that undermines the traditional assumption of epistemic transparency.

**Dedication**

To my parents, Evelyn and Nicholas Ciurria, and my sisters, Andrea, Julie, and Vanessa.

**Acknowledgements**

In this dissertation, I defend a situationist, externalist account of moral responsibility, which follows from research in situationist psychology. I do not discuss character, which is the central locus of situationist scholarship, but my view, following Maria Merritt, is that the most effective method of achieving regularly recurring behavioural dispositions—to the extent that this is possible—is to foster reliable social relationships and settings. However, in addition to this, one also requires a great deal of good luck. I am happy to say that I have had an abundance of both. Many people have contributed to the development of this dissertation, by reading drafts, discussing ideas, offering feedback, and, most importantly, providing moral support. I hope that I can return their generosity by reciprocating their kindness and emulating their example.

First, I want to thank the York University Philosophy Department faculty, students, and staff. In particular, I owe a debt of gratitude to my supervisor, Robert (Bob) Myers, and my second and third readers, Alice MacLachlan and Michael Gilbert. Bob generously read drafts of my chapters and gave invaluable feedback; Alice challenged me to work harder than I would have if left to my own devices; and Michael introduced me to coalescent argumentation, which gave me a better appreciation of the interpersonal dimension of moral reasoning. I am also indebted to Michael Guidice who, out of sheer altruism, vetted several of my drafts, and Duff Waring, who gave inspiring lectures in philosophy of psychiatry. I owe thanks to Andrew Sneddon for providing tough but fair criticism to what I expected to be the final draft of my dissertation, prompting substantial revisions. And I am grateful to my fellow graduate students, especially Khameiel Altamimi and Stephane Savoie, whose friendship and support have been invaluable to me.

There is probably not a single philosopher who has been more influential to me than John Doris, whose work is cited in almost every chapter of this dissertation. I had the honour of working with him after he responded to my impromptu e-mail—signed from an anonymous PhD student from Toronto—and had enough faith in me to phone me from Washington University to discuss my work. To my delight, he accepted a position on my supervisory committee, and has been a source of inspiration ever since. In addition, I owe thanks to John Faraci for providing feedback on my fourth chapter and supplying an advance copy of his forthcoming publication on responsibility, and Heather Battaly, who also provided me with an advance copy of her work. Finally, I am indebted to the many anonymous reviewers from various journals who commented on drafts of these chapters, and helped me to improve substantially on the originals.

Last but certainly not least, I owe thanks to my father Nicholas Ciurria, my mother Evelyn, and my three younger sisters, Andrea, Julie, and Vanessa. As the formative influence in my life, my family is largely responsible for both my abiding interest in ethics, and the development of my ideas on moral responsibility. They also helped me more directly by

discussing my chapter ideas at length, commenting on preliminary drafts, and allowing me to rehearse all of my presentations on them, even if the content sometimes exceeded their range of expertise. Their unfailing support has, to my mind, perfectly exemplified the moral of situationism in my view, i.e., that interpersonal relationships are the key to a flourishing life.

**Table of Contents**

**Chapter 1: Introduction**

Situationism is the view that there is no such thing as broad, situation-invariant character. This position emerged from experiments in situationist psychology conducted from the 1920s to the present, including the famous Milgram Obedience Experiments (1960-64), the Asch Conformity Experiments (1951-56), and the Darley and Batson Good Samaritan Study (1973). These experiments will be described in coming chapters, with special emphasis on the Milgram experiment (see chapter 2.3 for a preview,) which has been described by scholars, without exaggeration, as the most famous social psychology experiment of all time. This was the description given to me, for instance, by a speaker at the recent Obedience to Authority Conference at Nipissing University, Ontario (2013), which drew a multiplicity of scholars ranging from philosophers to social scientists to Milgram's original colleagues. The remarkable thing about these experiments is that they have continued to capture the popular imagination for the last 93 odd years, recurring in popular formats such as ABC's television program 'What Would You Do?,' Discovery Channel's 'Curiosity: The Milgram Experiment,' and MTV's 'The Challenge.' Milgram's experiments were originally conducted to determine whether any ordinary human, regardless of socioeconomic situation, would submit to Nazi authority in 1940s-era Germany, or if there was something distinctly evil, or at least perversely conformist, about the German psyche. Of course, the result was that most subjects submitted to the authority, and 'shocked' the victim to the hypothetical point of death, which seems to confirm Hannah Arendt's (1963) philosophy that evil is a banal phenomenon, to which ordinary people are highly susceptible. It is nothing to do with the German character *per se*. Perhaps these studies continue to hold people's attention because they provoke apprehension about our own character. They force us to ask ourselves such questions as, 'Could I turn out to be evil?' 'How well do I know myself?' and 'Do I even have a self?'

The situationist's response to the first two questions is, 'Yes, you could turn out to be evil,' and 'No, you don't know yourself particularly well.' The third question is more complicated, and most situationists, including Doris and Zimbardo, would say that you do have a self, but it is a very precarious, situation-sensitive self—so precarious, indeed, that you may be surprised to find that you are wont to go against your better judgment when you encounter seemingly mild countervailing situational pressures, such as an authority figure in a white lab coat telling you, 'the experiment must go on,' even though the parameters of the experiment no longer make logical sense. (What is the purpose of continuing to shock someone who seems to be unconscious, and how on earth could this test learning?) These may not be the answers that one wants to hear, but they are, to paraphrase Al Gore, inconvenient truths. We might want to think that we have robust character, that we are internally psychologically coherent, and that we have good introspective access to our intentional states, as Descartes and other 17$^{th}$ century rationalists would have us

believe, but endorsing this epistemological thesis may actually harm us. For instance, it could lead us to brazenly enter situations that we would be better off avoiding, or to make irrational judgments without considering all the relevant factors—factors that may be imperceptible to the untrained eye, and obscure even to experts (viz., Kahneman 2011). Recently there has been a wellspring of publications on how we can use the lessons of situationism to enhance our normative competence, including Francesca Gino's 'Sidetracked: Why Our Decisions Get Derailed and What We Can Do to Stick to the Plan' (2013), Michael S. Gazzaniga's 'The Ethical Brain: The Science of Our Moral Dilemmas' (2005), and Daniel Kahneman's 'Thinking, Fast and Slow' (2011). These all make valuable contributions to situationism, and inform the philosophy of moral education. While I want to give credit to this valuable interdisciplinary research, it is not the central focus of my dissertation.

My focus is moral responsibility, which I define, following Strawson, as deployment of the reactive attitudes, such as resentment, indignation, disapprobation, and indignation. In particular, I focus on negative moral responsibility, consisting of the negative reactive attitudes, especially blame. Now, I think that we can agree that situationists get something right. We might not be convinced that there is "no such thing as character," as Harman provocatively proclaimed (1999, p. 328), but there are degrees of situationism, which is conspicuous in the literature, as conveyed in my second chapter which focuses on the three preeminent situationists, Harman, Doris, and Zimbardo. Many people are not convinced that there is no such thing as character, but this does not mean that they cannot endorse the more modest situationist claim that character is so contingent upon situational factors that a psychologically normal person, who is an outstanding moral agent in ordinary life, might be induced to shock a supposed victim past the expected threshold of 150 volts (determined by Milgram through surveys) under fairly moderate inducements. Granting this fairly limited, and fairly well-supported, thesis, what implications can we draw? One of the main implications drawn by Doris is that we do not have good introspective access into our psychological states. He observes that, "given the well-known (and well-documented) schisms between what we say we care about and what we do, testimony about oneself is often unreliable testimony" (2007, p. 525), which implies that human psychology is fairly opaque, especially from a second-hand perspective. (If we cannot reliably access our own states, what chance do we have of accurately ascertaining other people's?) Doris refers to this practical difficulty as 'The Epistemological Problem,' and it is a key argument in my dissertation. If the epistemological problem is real, this means not only that character as a psychological construct is difficult to define and attribute across contexts, but also that second-personal psychological accounts of moral responsibility—which, it so happens, are the dominant approach in moral philosophy—are going to face a major empirical hurdle, inasmuch as they depend on the tangibility of second-hand epistemic data. It also implies that while it is seductive to construct hypothetical scenarios in which, say, Smith intends to shoot Jones for financial gain, if the human

psyche is fundamentally obscure, then these thought experiments actually are not very helpful. I suggest that a range of quantifiable data is required to assess a person's responsibility status, meaning that we must know much more about Smith, Smith's circumstances, Smith's personal history, and general facts about human cognition (such as types of cognitive bias).

Adding to this difficulty, neuroscientific studies corroborate that the mind is murkier than Descartes presumed, owing to the pervasiveness of such cognitive distortions as self-serving bias, confabulation, and rationalization. I highlight some of these distortions in my second chapter, citing studies by Bejamin Libet, Gazzaniga, and Zimbardo. These epistemological issues call into question the viability of the traditional internalist approach, and give us reason to search for an alternative account of responsibility that does not depend on introspective/psychological evidence. It just so happens that such an account is readily available, and has been in use for at least the last 132 years in legal philosophy, in O.W. Holmes' venerable, 'The Common Law' (1881). This account is known as *the standard of the reasonable person* [SRP], which measures responsibility according to what a reasonable person, properly conceived, would do in the subject's circumstances. From an empirical perspective, in light of situationist and neuroscientific research, this view has clear advantages over the internalist method. SRP measures responsibility according to what a reasonable person would do in the target agent's circumstances, meaning that it compares an actual agent's overt behaviour against a hypothetical agent's overt behaviour, thus dispensing with subjective data. SRP is particularly useful for dealing with cases of negligence, coercion, compulsion, and necessity, which impute culpability regardless of the agent's subjective intentions; but I believe that it can be extended to all instances of moral responsibility ascription if we construe it in terms of *avoidability*, defined as a function of the agent's action options, personal history, and epistemic circumstances. The key question then becomes, 'What could the agent have done in the circumstances, given the situational possibilities and epistemological resources available to her at the time of action, and those available to her during the course of her personal development?' This approach shifts the focus from the agent's psyche to her present and past environment, and uses testable empirical data to gauge her responsibility. I flesh out this account in my first three chapters.

I have defined the alternative to the 'externalist' method as 'internalist,' but what exactly is internalism, and what theories in moral responsibility can be classified under this heading? By internalist, I mean any moral theory that measures responsibility *exclusively* in terms of the agent's psychological states. The same demarcation is made in the philosophy of law under the rubric, 'objective' and 'subjective,'[1] with subjective accounts comprising two main categories: choice theory and character theory. I shall leave the details of this taxonomy to chapter 4, where I outline the analogues of my view in legal philosophy. For now, it suffices to say that internalist

[1] These are well-known views within legal philosophy. I use the terms 'internalist' and 'externalist' as opposed to 'subjective' and 'objective' when referring to theories of moral responsibility, to

theories, on my conception, are ones that ascribe responsibility on the basis of the agent's psychological states, particularly her choices and character traits (settled patterns of thought and action). I may be the only scholar to have drawn the internal-external distinction in moral philosophy, aside from Andrew Sneddon (2005), who seems to have coined the term 'internalism' as a label for psychology-based moral theories. I came across Sneddon's work after having completed much of my thesis, and was surprised to find that someone else had presented a roughly analogous account. It was comforting to find that someone else had made inroads in my unique field of inquiry, and it brought to mind Peirce's view that, "the remarkable convergence of ideas of two thinkers… adds an extrinsic confirmation" to the view (Potter 1996, p. 124). While Sneddon has developed externalism as a philosophical methodology, he has only, to my knowledge, published one paper exclusively on externalism concerning moral responsibility (2005), and part of a chapter from his book (2006). In spite of similar methodological principles, our arguments are quite different, but, I believe, mutually reinforcing. After delineating my position for five chapters, I adduce Sneddon's externalistic interpretation of Strawsonian responsibility in chapter 6. I present new, supporting arguments for Sneddon's view, and also highlight a hitherto-unnoticed implication of the externalist approach, viz., that it justifies Strawson's famous excuse of 'peculiarly unfortunate formative circumstances.' If my analysis is correct, it quells a longstanding debate about this excuse, perpetuated by Watson (1986) and Scanlon (1987), amongst others. It also connects philosophical inquiry with research on the psychology of excuses, particularly Faraci and Shoemaker's (2013) investigation into naïve moral reasoners' intuitions about deprived childhood circumstances, which will be discussed in chapter 7. There, I analyze the researchers' results and dispute their conclusions, which deny the moral relevance of deprived childhoods. I argue that the most intuitive reading of the results supports a *pro tanto* excuse on these grounds. That is, I argue that if someone's circumstances were especially epistemologically constraining, the agent is *to that extent* excusable. This viewpoint is, I contend, borne out by the authors' survey results, and also introduces a scalar aspect to moral responsibility assessment. These arguments help to settle some unresolved questions in moral philosophy, but they also point to the need for more empirical research, which I hope to undertake in the near future.

Now, which theories of moral responsibility can be called *internalist* on my definition? In my opinion: all of them. Or at least, all those in analytic philosophy. (There may be comparable approaches to mine in continental philosophy, and I have noticed some analogues in Merleau-Pontian phenomenology and Heideggerian hermeneutics, but I am not sufficiently well versed to say whether these are clear, coherent, and internally consistent. My suspicion is that there are not.) Within analytic philosophy, at least, with the exception of Sneddon, I take myself to be the only philosopher offering an explicitly externalist account of moral responsibility. There are, no doubt, hints of externalism in the literature. For example, Susan Wolf (1987) describes deprived

childhood circumstances as excusing, J.M. Fischer (1987) cites historical circumstances as morally relevant, and Pettit and Smith (1996) describe responsible beliefs as consistent with "the evidence," which is supposed to exist objectively in the world (p. 403). Similarly, Miranda Fricker (2007) describes culturally-induced ignorance as a potential excusing factor. But these authors at times seem to want to reduce these conditions to internalistic criteria, such as defects or incongruities in the agent's psychological economy. Nonetheless, there are tensions in their works that suggest that such a reduction is impossible, or at least in conflict with illustrative examples that they give, which seem to favour externalism. If these philosophers intend to present an externalist account of responsibility—which is difficult to discern—they nowhere attempt to defend this methodological picture. But nor do they try to defend an internalist approach. They are simply silent on the question of method, which suggests that they have not even considered which framework they mean to endorse. This lack of insight leads to confusion, vagueness, and inconsistency in moral philosophy. Moreover, inasmuch as the methodological question is ignored, all moral philosophers have, to this point, implicitly begged the question. By defending an externalistic theory of responsibility, I bring this methodological bias to the fore, and force a discussion of the merits of the two approaches. Philosophers are henceforth compelled to defend a particular metaethical picture.

I am not sure how my dissertation will affect general opinion about situationism as an ethical paradigm. One thing that I have noticed is that there is a stark contrast between popular reception of situationism, which has been generally very favourable (as exemplified in the numerous television replications of the experiments,) and scholarly reception, which has been less enthusiasic. Unlike laypeople, many moral philosophers actually appear to hate situationism. Not just to dislike it, but to hate it. This has been more conspicuous to me in personal encounters than in academic writing, where philosophers tend to take a detached, dispassionate stance. By contrast, when I was asked about my specialization at a philosophy party recently, the inquirer scoffed, 'Oh, situationism, isn't that the view that because some people found dimes in phone booths, there's no such as character?' (This comment refers to the Isen and Levin Phone Booth Experiment [1972] in which some subjects found a dime that had been left in a phone booth by the experimenters, and others did not, and those who found the dime were 22 times more likely to help a 'victim' in a subsequent test case.) At another conference, an acquaintance described situationism as a 'nihilistic theory with no real-world application.' I remember these incidents distinctly, as they made me wonder why situationism is met with such hostility and antagonism compared to other theories. My suspicion is that many philosophers are confused about what situationism means. If they have only a cursory familiarity with the literature, they may think that situationism is supposed to imply that there is no such thing as character in any sense of the word, and thus there can be no use for either ethics or moral education—that is, ethicists should

be out of the job. But this certainly is not the standard view. (Harman is probably the only one to defend such a deflationary reading, but even he believes in culturally relative moral norms, so falls short of endorsing moral nihilism.) In addition, virtue ethicists may resent Doris' 1998 contention that situationism refutes virtue ethics; but they should equally resent Kantianism and consequentialism, which are equally logically disqualifying (via the principle of non-contradiction.) For my part, I actually endorse a version of virtue ethics, but it happens to be incompatible with a certain reading of Aristotelianism (which posits motivationally self-sufficient virtues), as well as Kantianism and consequentialism generally. This means that I am in a position to make many enemies, but surely not to a greater extent than any other moral philosopher. So it remains unclear why situationism is not well-received.

I can only speculate as to the reason, but it may be that people misunderstand the implications of situationism, in which case this dissertation should foster greater acceptance. One brief clarification that I shall make here, as a sort of prologue, is that situationism shares strong affinities with feminist philosophy and virtue epistemology. I can only adumbrate these similarities here, but they will gradually come into focus. In feminist philosophy, there is a common tendency to explain unethical behavior such as identity prejudice in terms of structural factors as opposed to, or at least to a greater extent than, agent-centric variables; and in virtue epistemology, there is a tendency to focus on an agent's epistemic environment, as opposed to, or more than, the agent's subjective states. Examples of this orientation include feminist social epistemologists such as Elizabeth Anderson 1995 and Miranda Fricker 2007, and Marxist feminists such as Martha Gimenez 1990 and Teresa Ebert 1996. These theories sometimes hold agents responsible for their ignorant and prejudiced behaviour, particularly when this behaviour is beyond the pale; but they often shift blame onto systemic factors that preempt large-scale social currents, of which individual agents are merely a component. Many feminists believe that this broader structural emphasis is effective at targeting identity prejudice at its roots, as opposed to singling out individual instances instantiated in the actions of discrete human subjects, who may, in any case, simply be victims of '*epistemic bad luck'* (to borrow a phrase from Fricker 2007, p. 101). The former, systemic approach is not only (arguably) more efficacious at opposing adverse social trends, but fairer to individual humans who may not be able to fathom a more enlightened perspective from their embodied social location. The structural view also helps to deflect blame from women who are complicit in patriarchal structures due to oppressive patriarchal forces, not personal failings.

This is admittedly a very cursory overview of what I take to be important similarities between situationism and feminist thought. They will become increasing perspicuous in what follows, and they likely warrant a book-length treatment in their own right. I mention them here because I believe that this comparison may shed light on situationism's more favourable application, as a natural complement to feminist philosophy. This contradicts what many

philosophers apparently think of situationism, viz., that it is a misguided, nihilistic attack on moral ideals, especially moral realism. Situationism may seem heretical—and to some extent *really is* heretical—because, like feminist philosophy, it calls into question the traditional agent-centric view of the person that has dominated western philosophy since at least 1641, when Descartes introduced the figure of the autonomous, epistemically transparent rational thinker. Feminist epistemologists, such as Helen Longino (2002) and my departmental colleague Lorraine Code (1991), have rightly disputed this atomistic conception of the knowing agent, and situationism, I believe, carries on this legacy by drawing attention to our social embeddedness and the influence that social relations and setting can have on human behaviour. This means that situationism is not the cynical, inflammatory view that some people perceive, but rather a tried-and-true methodology that illuminates and empirically corroborates longstanding intuitions in feminist thought. Much research in situationism has been concerned to investigate the relationship between such distorting factors as attribution error, self-serving bias, and social conformity bias (on one hand,) and gender prejudice (on the other.) Thus, situationism is already, at least implicitly, yoked to the practical concerns of feminism. I believe that this coalescence will only increase as situationists focus increasingly on epistemic responsibility, and its basis in human cognition.

To briefly summarize my dissertation, the second chapter (following the introduction) highlights the inconsistencies and logical incongruities in the situationist literature, and offers a solution in the form of an externalist approach, as embodied in SRP. The third chapter sheds light on how to understand and apply SRP appropriately in light of the legitimate challenges lodged by feminist philosophers, critical race theorists, and critical disability scholars. The fourth chapter explicates the relationship between legal and moral responsibility, which diverges somewhat from the commonsense understanding. The fifth chapter addresses an ostensible difficulty for SRP: judging responsibility for actions stemming from mental health disabilities, which are internal in nature. I argue that SRP is not only adept for such cases, but preferable to character-based alternatives. In my sixth and seventh chapters, I shift gears somewhat and focus on externalism as method, provide a defense of this approach to responsibility, and draw implications for understanding the psychology of excuses. This lends ballast to moral-SRP, and situates it in the taxonomy of ethical theories alongside Strawson's account of moral responsibility. I use Sneddon's interpretation of Strawson to support my account, and then I defend the excuse of deprived childhood circumstances. In chapter 7, I continue this line of argument, contending that externalistic excusing factors are deeply entrenched in our shared pretheoretical intuitions. This argument hinges on my interpretation of research conducted by Faraci and Shoemaker into folk psychology.

## Chapter 2: Situationism and Negative Moral Responsibility

### 1. Preface

This first substantive chapter develops and defends a situationist account of moral responsibility. The starting point for this position is the realization that situationists defend widely varying accounts of moral responsibility, ranging from radically revisionary (viz., Harman) to fairly conservative (viz., Zimbardo,) and which include views that appear to be incongruous with the central logic of situationist psychology (viz., Doris.) Doris defends an identificationist, Frankfurtian account of responsibility, which defines responsibility in terms of one's ability to form second-order volitions, i.e., higher-order desires that comprise executive cognitive function, and are unique to human beings. These volitions are also defined by Frankfurt as the core of character. But since situationism entails that there is no such thing as character in this sense, Frankfurt's view cannot possibly provide a coherent foundation for a situationist theory of responsibility. (If we tried to define the agent's deep self in narrow characterological terms, this would seem to contravene the spirit of Frankfurt's account.) As a result, I identify a need for a situationist-compatible account of responsibility, and offer such an account in this chapter. I propose that we need an *externalist* (as opposed to internalist) standard of responsibility, which dispenses with the psychological states of the agent, as provided by introspection—which empirical research reveals to be opaque, unreliable, and fragmented. Externalism, by contrast, emphasizes situational variables such as the action options available to S, S's personal history, S's present environment, relevant base rate statistics, and relevant cognitive science data. Given that such a standard already exists in legal philosophy in SRP, I submit that moral philosophers should adopt a version of this standard for assessing moral responsibility. Moral-SRP (as I shall call it) offers a convenient ingress for situationist theorizing about responsibility, and is already justified in moral terms in Holmes' defense (via principles of fairness and social welfare.) Thus, it is, in essence, already a moral viewpoint. However, I acknowledge that we need to refine SRP, which was originally inauspiciously designated 'the standard of the reasonable man,' to accommodate the concerns of feminists, critical race theorists, critical disability scholars, and other interest groups. In the third chapter, I continue this line of inquiry by suggesting ways of refining SRP more precisely, though much of this work must necessarily be left to informed discretion.

Before proceeding with the situationist analysis, however, an account of moral responsibility is in order. Specifically, what is it?

### 2. Moral Responsibility

I take it that moral responsibility is a fairly intuitive concept. If someone fails to return a library book without good reason, the person is morally responsible. More iniquitously, if someone wantonly tortures puppies, the person is presumptively morally responsible. These examples of responsibility are negative, and negative cases are the central focus of this treatise. I suspect that there are asymmetries between negative responsibility (blame, paradigmatically) and positive responsibility (praise, paradigmatically,) but I will not explore them here, as elucidating the distinctive nature of praise would require another 60,000-word volume. At any rate, negative responsibility is an important concept in its own right and warrants extended analysis. Now, although our intuitions provide a fairly reliable barometer of negative moral responsibility, and it is easy to envision myriad uncontroversial cases of responsibility ascription, it is helpful and illuminating to bolster this concept with an account from moral philosophy. My view of moral responsibility is closest to Strawson's account of responsibility as deployment of the 'participant reactive attitudes,' which consist (in their negative form) of blame, resentment, disapprobation, indignation, and the like. The reactive attitudes, according to Strawson, are natural interpersonal reactions to other people's behaviour. I will discuss this account in some depth in chapter 6, but for now it suffices to give a preliminary sketch. Strawson believes that the reactive attitudes are justified by our unreflective interpersonal practices, but philosophers such as Watson and Scanlon have pointed out that a normative rationale is needed. I will suggest such a rationale shortly, but first it is important to understand Strawson's central argument. He identifies two types of excusing conditions: (1) The agent "didn't mean to," "didn't know," "couldn't help it" (p. 77), and (2) The agent "wasn't himself," "has been under very great strain," is "systematically perverted," or suffered from "peculiarly unfortunate formative circumstances" (pp. 78-79). The boundary between the two types of excuses is supposed to be permeable, but the second type is more excusing (and thus, sometimes called 'exempting'). In what follows, I will argue that type-1 and type-2 excusing conditions are based on the rationale that the agent was not reasonably able to avoid a given transgression, in accordance with SRP. That is, I will show that Strawsonian responsibility is compatible with SRP, and that this explanatory foundation succeeds in accommodating challenges from situationist psychology. This account also respects Strawson's requirement that we justify the reactive attitudes from within the 'participant stance,' as it relies on justificatory principles that are intrinsic to commonsense morality.

**3. Situationist Psychology**

Below are some of the most-cited experiments in situationist psychology.

1. **The Milgram Obedience Experiments**

   In these experiments, conducted multiple times at Yale and Bridgeport, the subject was assigned to the role of teacher, and instructed to administer increasingly intense electrical shocks to a learner. The learner, however, was the experimenter's confederate, and the shocks were simulated. The subject was instructed to shock the learner each time a question was answered incorrectly, putatively to test the effects of punishment on learning. The subject received a shock machine with shocks ranging from 150 volts, labeled 'Slight Danger,' to 450 volts, labeled 'XXX' (just after 435 volts, 'Danger: Severe Shock.') As the subject shocked the learner, the latter demonstrated pain behaviours, including grunting at 75 volts, "shout[ing] at 120, ask[ing] to be let out at 150, scream[ing] in agony at 270, and go[ing] silent at 300" (Patten 1977, p. 426). In other experiments, the learner pounded on the door, asked to be let out, and complained of a heart condition just before falling silent (Milgram 1963). If the subject expressed concern, the experimenter would respond with one of four prompts: (1) 'Please continue' or 'please go on,' (2) 'The experiment requires that you continue,' (3) It is absolutely essential that you continue,' and (4) 'You have no other choice, you must go on.' The experimenter could say nothing else. Milgram and others whom he questioned predicted that very few people would surpass the threshold of 150 volts, designated 'Very Strong Shock.' However, in one characteristic experiment (1963), all of the subjects surpassed this threshold, and 65 percent continued to the 450-volts mark, at which a real subject would likely have been dead or at least severely injured. Milgram concludes, "something… dangerous is revealed: the capacity for man to abandon his humanity, indeed, the inevitability that he does so, as he merges his unique personality into larger institutional structures" (1974, p. 188).

2. **The Darley and Batson Good Samaritan Experiment**

   In this experiment, Princeton seminary students were recruited to give a talk at a building across campus. First they were asked about their motives for pursuing religious enlightenment. Then, half of them were assigned to read a passage on seminary students' vocational options, and half were assigned to read the Good Samaritan parable. Finally, some students were told that they were late for their presentation and should hurry, others were told that they had just enough time, and others were told that they were early. On the way to their destination, the subjects passed an ostensible victim hunched in a doorway, moaning in pain (who was in fact the experimenters' confederate.) The experimenters were interested in studying whether the seminarians' motives (the character variable,) the assigned story, or time constraints would influence helping

behaviour. They found that the only variable that made a significant difference was time. Only 10% of those in a hurry stopped, compared with 45% of those in a moderate hurry and 63% of those in no hurry. The experimenters concluded that, "only the hurry main effect was significantly . . . related to helping behavior" and "only hurry was a significant predictor of whether one will help or not." Subsequently, Harman (1999) further inferred that people systematically commit the fundamental attribution error, by (mis)attributing antisocial behaviour to character as opposed to situational factors.

3. **The Isen and Levin Phone Booth Experiment**

   Here, the experimenters planted a dime (the cost of a phone call in 1973) in some phone booths but not others. Then they placed a confederate outside of the phone booth, prepared to drop a folder full of papers in the path of emerging callers. Of those who found a dime, 14 helped and 2 did not, and of those who did not find a dime, 1 helped and 24 did not. It seemed that the subjects' elevated mood upon finding the dime disposed them to help the victim, regardless of their broader character traits. This leads Doris (1998) to conclude that, "the behavioral reliability expected on standard theoretical constructions of personality is not revealed in the systematic observation of behavior" (p. 504).

These are some of the experiments that support the conclusion that there is no such thing as situation-invariant character.

**4. Précis**

For the remainder of this chapter, I present a situationist account of negative moral responsibility.[2] In *Moral philosophy meets social psychology* (1993), Gilbert Harman argues that social psychology can educate folk morality to prevent us from committing the "fundamental attribution error" (FAE), i.e., "the error of ignoring situational factors and overconfidently assuming that distinctive behaviour or patterns of behaviour are due to an agent's distinctive character traits" (p. 315). If we survey the literature, we find that situationists agree with Harman regarding character, but they disagree about whether we also commit FAE concerning moral responsibility. Do we ignore situational factors and overconfidently assume that people are *morally responsible* for their negative behavior? Harman seems to think so, but other situationists (viz., Doris 2002 and Zimbardo 2007) evidently disagree. It is particularly difficult to adjudicate this question given

[2] For simplicity, I will use the terms 'negative moral responsibility,' 'moral responsibility,' and 'blame' interchangeably in what follows. I hope that the reader can ignore the minor differences amongst the three concepts.

that no situationist has, to my knowledge, conducted an adequate philosophical analysis of moral responsibility, and this creates a conceptual gap. I aim to show that situationist psychology impugns psychological accounts of moral responsibility, and supports SRP, due to a multiplicity of epistemological difficulties.

**5. Inconsistencies and Ambiguities**

Harman argues that social psychology can teach us to avoid "overconfidently assuming that distinctive behavior or patterns of behavior are due to an agent's distinctive character traits" (1999, p. 315). For present purposes, I am not interested in virtue ethics or character (see Ciurria 2014 for this account); rather, I am concerned to pursue Harman's interdisciplinary project in another way. Specifically, I intend to investigate the implications of social psychology for moral responsibility, to see if the research can provide guidance.

Most scholars have focused on the implications of social psychology for globalist conceptions of character (e.g., Staub 2004, Baumeister and Vohs 2004, Duntley and Buss 2004), but few have extended the inquiry to moral responsibility judgments. One notable exception is Arthur G. Miller et al. (2002), who argue that while social psychology explanations do not absolve perpetrators, they nonetheless "humanize them, and, by extension, [urge us to] indict ourselves as potential perpetrators" (p. 320). This suggests that such explanations may, in some sense, excuse, but the relevant sense remains somewhat mysterious. Another exception is Garrath Williams (2003), who remarks that, "there is clear evidence from social psychology that *blame* is especially frequently and inappropriately attributed to individuals in modern western society" (p. 427, emphasis mine.) This interpretation might seem to follow logically from FAE: just as we neglect situational factors and overeagerly attribute robust character to individuals, we must also, for the same reasons, overeagerly attribute moral responsibility. But this inference needs argumentative support. Moreover, an overview of the literature reveals that while situationists unanimously agree about the misattribution of robust character, they differ quite drastically on the subject of moral responsibility. This is partly because they have paid so little sustained attention to this important area of moral philosophy, perhaps mistakenly seeing it as too intuitive to warrant scrutiny, or immune from situational analysis. At any rate, their accounts show significant disparities. To illustrate, consider three of the most prominent situationists, Harman, Doris and Zimbardo:

1. Harman seems to agree with Williams. He writes, "when things go wrong, we typically *blame* the agent, attributing the bad results to the agent's bad character"; however, "a greater understanding of the agent's situation and how it contributed to the action can lead to a greater tolerance and understanding of others" (1999, p. 329, emphasis mine.)

This suggests that lack of character and lack of responsibility go hand in hand, and thus FAE applies equally to character and responsibility. Furthermore, Harman is a moral relativist, and his relativist account (1975) exonerates many people who would count as blameworthy on a folk morality view—for instance, criminals who grew up in a criminal family and internalized their family's anti-social values. So his account of moral responsibility is far more lenient than commonsense morality would allow.

2. Doris provides the most comprehensive philosophical account of moral responsibility, in a chapter of his book, 'Lack of Character' (2010). However, it is bizarrely incongruous with the central logic of situationist psychology. He defends an 'identificationist,' Frankfurtian account of moral responsibility, according to which a person is responsible only for actions that she endorses at the level of her higher-order volitions or deepest self. However, Frankfurt defines a person's second-order volitions as the essence of her deep-seated character, and Doris is opposed to character in this sense. Hence, without further argument, it is unclear how this view is supposed to be compatible with situationism. This issue will become clearer in what follows, as we see that Doris believes that human psychology is epistemically murky and inaccessible to introspection. In his 2010 article, he admits that it may be difficult to discern whether an agent did in fact endorse his action, but he dismisses this worry by saying that, "in such instances ambivalence is the appropriate judgment—or abstention from judgment" (p. 145). However, if the epistemological problem is as serious as he submits in 2006, then it seems that abstention from judgment will be ubiquitous: we will virtually never be able to judge whether a person is morally responsible.

3. Zimbardo provides a third perspective. He repeats many times that he is not in the business of "excusiology": "Individuals and groups who behave immorally or illegally must still be held responsible and legally accountable for their complicity and crimes"; however, "in determining the severity of their sentence, the situation and systematic factors that caused their behavior must be taken into account" (2007, p. 231). Zimbardo believes that Western society should shift its focus from an "individualistic orientation," which regards "the person as sinner, culpable, afflicted, insane or irrational," to a systems analysis, which emphasizes systems and situations (2007, p. viii). This seems like a reasonable position for a situationist to hold, but Zimbardo does not provide any justifying rationale. Moreover, when he contends that one of the 'guards' in the Stanford Prison Experiment (SPE), Hellman, was distinctively blameworthy because he went "beyond the demands of the situation to create his own 'little experiment' to satisfy his personal curiosity and amusement" (2007, p. 218), he provides no reason to draw a distinction between

> Hellman and the others. Thus, while Zimbardo's intuitions seem right, a philosophical defense is exigent.

The main issue is that these three situationists provide three disparate accounts of moral responsibility, and only Doris even attempts a philosophical analysis. Unfortunately, this analysis is logically incoherent. Due to this lack of sustained reflective attention, situationists tend to rely heavily on brute intuitions. For example, Harman's argument ostensibly rests on nothing more than a rhetorical question about participant base rates:

> Can we really attribute a 2 to 1 majority response to a character defect? And what about the fact that all subjects were willing to go at least to the 300 volt level? Does everyone have this character defect? Is that really the right way to explain Millgram's results? (Harman 1999, p. 322)

Harman nowhere tries to justify this intuition, and virtue ethicists seem to have opposing inclinations (e.g., Rachana Kamtekar 2004). Later in the same article, Harman asserts that it should also seem odd to blame a violent guerilla in cases of widespread violence such as the Yugoslav wars, but never philosophically unpacks this claim (p. 329). This is a setup for a battle of intuitions. In the next section, I hope to give some argumentative heft to Harman's intuitions about responsibility in the Milgram experiment, using an apt standard from legal jurisprudence. Specifically, I will appeal to SRP, which takes into account the agent's capacities and relevant features of the situation. SRP also justifies Zimbardo's view that systemic factors can excuse blame.

**6. The Standard of the Reasonable Person (SRP)**

I think that there is something to the intuition that the Milgram participants were not (fully) blameworthy, given that all of them surpassed the expected threshold of 150 volts, and most of them far surpassed it, even though they were psychologically normal adults. Of course, this does not mean that they behaved rightly, or that their actions were permissible; what they did was morally wrong, but excusable. (Excusability in its very nature implies the commission of an objective infraction—otherwise there would be nothing to excuse.) The difficulty for the philosopher lies in explicating what underscores this intuition. I believe that this problem can be resolved by appealing to SRP, as a moral heuristic. In its original, legal form, SRP defines responsibility, in Holmes' words, according to "what would be blameworthy in the average man, the man of ordinary intelligence and prudence" (1881, p. 73). In contemporary scholarship, SRP has been expunged of its unsavory masculinist connotations, and adjusted so as to

accommodate the interests of a broad range of uniquely-situated political identities. For example, in the landmark case, R. v. Lavallee (1990), the court acquitted Lyn Lavallee of murdering her common-law husband after recognizing 'battered woman syndrome' as a legal defense. The court argued:

> If it strains credulity to imagine what the "ordinary man" would do in the position of a battered spouse, it is probably because men do not typically find themselves in that situation. Some women do, however. The definition of what is reasonable must be adapted to circumstances which are, by and large, foreign to the world inhabited by the hypothetical 'reasonable man' (Bickenbach 2007, p. 277).

Other modifications have been proposed to accommodate children, persons with disabilities, and political minorities (see Moran 2003). I address these concerns in chapter 3, but for now I will restrict deliberation to straightforward cases of psychologically and circumstantially normal adults. Penumbral cases will be discussed later, but a degree of discretion must be exercised on a case-by-case basis, as with any principle. For present purposes, my aim is very general—namely, to show that SRP has *pro tanto* extenuating implications for normal adults, when the case admitted of a clear excuse (e.g., duress, coercion, and the like).

It is noteworthy that Holmes' defense of SRP has both inculpating and exculpating implications. On the one hand, if a person is "born hasty and awkward" and is "always having accidents," his neighbours and the courts can require him "to come up to their standard" (p. 73). But on the other hand, "when a man has a distinct defect of such a nature that all can recognize it as making certain precautions impossible, he will not be held answerable for not taking them" (p. 74). The rationale for this proposal is twofold. First, "the impossibility of nicely measuring a man's powers and limitations" impugns psychological evaluation (1881, p. 74); and secondly, utilitarian considerations, which are traditionally paramount in legal theory, must be balanced against individual rights, to avoid treating people as a mere means. Accordingly, "when [people] live in society, a certain average of conduct, a sacrifice of individual peculiarities going beyond a certain point, is necessary for the general welfare," and yet "a blind [person] is not required to see at his peril" (p. 46).

In the following two sections, I analyze and enlarge upon each of these considerations, which I label (i) the epistemological problem, and (ii) the principle of reflective equilibrium. The first point states that we cannot, with sufficient reliability for the purposes of assigning punitive sanctions (either legal or moral,) assess an agent's motives or intentions, and thus we cannot gauge responsibility on this basis; and the second point states that we should, to the greatest extent possible, aim to protect the general welfare, while *pari passu* abstaining from blaming people for outcomes which they could not have *avoided,* in a reasonable, pragmatic sense of the

word. This sense will be explicated in sections 7 and 8, via Fricker's seminal account of epistemic (in)justice, which includes an epistemic reading of the familiar principle of ought-implies-can.

In chapter 4, I shall argue that there is a very close relationship between moral responsibility and legal responsibility, although there is (obviously) a significant difference between legal permissibility and moral permissibility. That is, while moral and legal *theory* are disjunctive (such that adultery, for instance, is a moral but not a legal infraction,) moral and legal *responsibility* are basically isomorphic on my conception. This is because both conceptions hinge on the agent's all-things-considered ability to conform to putative normative principles (of permissibility,) *whatever they happen to be*. I shall proceed on this assumption, but if the reader cannot accept this relationship provisionally, he or she is encouraged to advance immediately to chapter 4.

**7. The Epistemological Problem**

In an incisive critique of the two standard accounts of criminal liability—choice theory and character theory—Doris (2007) argues that both are inadequate due to their reliance upon the notion of 'settled character,' which is empirically untenable. *Choice theory* entails that someone is legally responsible if "he broke the law by an action which was the outcome of his free choice" (p. 521), and *character theory* entails that someone is legally liable if his action was the result of his settled character, i.e., settled patterns of thoughts and action. Doris rejects the first view because actions committed under duress or coercion may appear to be choices, in the minimal sense that they were committed voluntarily and according to the agent's wishes, but if they were performed under extreme unlawful pressure they preclude responsibility. Legal scholars have sought to resolve this problem by defining 'free choice' as an expression of one's "rational judgment," "evaluative commitments" or "plan of life," but on this description, choice theory "is slouching into character theory" (Doris 2007, p. 521). According to character theory, if a person has violated the law, then he is presumptively responsible; however, if he was acting 'out of character'—under duress, for instance—then he may be excused to some extent (ascertained by judicial discretion.) Doris objects to this view on two scores: (1) situationist psychology shows that there is no such thing as settled character, and (2) there is no reliable means of differentiating 'in-character' behavior, for which the agent is putatively responsible, from 'out-of-character' behavior, for which the agent is putatively excusable. At present, I am not interested in (1), because even Doris admits that there may be valid local conceptions of character, which, in principle, could suffice to establish responsibility for all intents and purposes. However, I believe that (2), which Doris terms 'the epistemological problem,' is especially vexing for character theory, and warrants especially incisive scrutiny.

Doris describes the epistemological problem as the problem of ascertaining, with

sufficient reliability for imposing punitive measures, whether the defendant was acting 'in character' or 'out of character' when he committed the act in question. How can this discrimination be made? There are two standard interpretations. We can explain 'in character' (a) statistically, or (b) in terms of frequency. The first interpretation fails because "many criminal actions, such as homicide, are relatively rare" (Doris 2007, p. 525), and yet this does not prevent us from sentencing first-time murderers; and the second interpretation fails because even if an act is "seldom undertaken relative to available opportunities," we do not wait until someone is a "career recidivist" to press charges (Doris 2007, p. 525). So neither interpretation captures our actual practice of assigning responsibility. Another layer of difficulty lies in the fact that testimony is often unreliable, especially when one's civil liberties are at stake, and there are countless omnipresent incentives for deceit. Thus, in addition to the problem of ascertaining character in terms of statistical regularity and frequency, it is exceedingly difficult to judge whether or not a person's testimony is trustworthy. Indeed, multiple studies find that people are no better than chance at detecting lying, and police officers with special training are no exception (Ekman & O'Sullivan 1991, Granhag & Strömwall 2004, Mann & Vrij 2004.)

These are familiar reasons for repudiating choice and character theory, but there is another type of difficulty which is seldom invoked in debates about responsibility, though it is an extension of the first. The problem is that we have difficulty discerning not only other people's motives, but also our own—and perhaps to an even greater extent, since we are subject to multiple cognitive distortions (viz., Kahneman 2011 for an extensive account). In other words, we have very weak, very unreliable introspective insight. This claim may seem counterintuitive to anyone who still subscribes to the Cartesian conception of the mind—which retains influence in analytic epistemology—according to which our inner states are introspectively available to consciousness. This view has been challenged by Freud's theory of the unconscious mind, and more recently by cognitive science research which calls into question both our psychological coherence and epistemic transparency. These studies present a picture of the mind as much more fragmented, inscrutable, and situation-sensitive than analytic philosophy suggests, or than phenomenological intuitions may reveal. The following three examples are paradigmatic of this research:

1. In tests on split-brain patients—in which the patient's brain is severed along the corpus callosum to prevent the spread of epileptic seizures—Sperry (1961) and Gazzaniga, LeDoux and Wilson (1977) found that "the left brain tended to confabulate about actions performed by the left hand, which is controlled by the right brain" (as reported by Hirstein 2005, p. 153). In one experiment, Gazzaniga et al. flashed commands to the patient's right hemisphere, which responds to visual inputs. Subsequently, they asked the patient why he was acting as he was, which was interpreted by the left hemisphere's verbal

processing mechanisms. “In trial after trial, when queried, the left hemisphere proved immediately adept at immediately attributing cause to the action” (Hirstein 2005, p.153). For instance, when ‘walk’ was flashed to the right hemisphere, and then the patient was asked why he was walking, in one typical experiment he replied ‘Oh, I need to get a drink.’ Patients consistently confabulated in this manner, prompting Hirstein to surmise that, “people might be doing this often in their everyday life… It makes some writers wonder whether we might all be doing something like this” (p. 154). This illustrates our potentially massive susceptibility to confabulation.

2. While conducting research into neural activity and sensation thresholds, Benjamin Libet (1992) discovered that unconsciousness neural processes in the brain called ‘readiness potential’ precede conscious decisions to perform volitional actions, implying (Libet argues) that these unconscious neural processes cause behaviors which the agent falsely experiences as consciously chosen. In one experiment, Libet asked his subjects to flick their wrist within a timed interval and report the instant when they were first aware of the intention to flick, while simultaneously pressing a button. He found a consistent latency between the patients’ readiness potential (i.e., the time of decision indicated by neuronal activity on the EEG) and reported time of decision, of 300-500 milliseconds. Like the split-brain experiments, these results suggest that our motives may be confabulated.

3. In another set of experiments, Gazzaniga (1992) found that people hardly ever admit to having made irrational decisions. In an experiment in which women were asked to choose a pair of nylon stockings from an assortment, and then explain their choice, all of them gave detailed and sensible reasons, not realizing that the stockings were in fact identical! That is, their ‘reasons’ were rationalizations. This can have dire consequences in morally dangerous situations, where people tend to rationalize their harmful actions. In SPE, for instance, Zimbardo found that when ‘guards’ were instructed to play a role that went against their better judgment, they universally provided “reasons why they were doing something contrary to what they really believed and what they stood for morally” (2007, p. 220). This prompts Zimbardo to surmise, “people are less rational than they are adept at rationalizing—explaining away discrepancies between their private morality and actions contrary to it” (2007, p. 220). Thus, there is cause to doubt our sincerest narratives about ourselves.

In the realm of *legal philosophy*, Doris takes the epistemological problem to entail that we should eschew choice and character theories in favour of an ‘objective’ theory of responsibility, SRP.

This standard avoids epistemological difficulties because, in Doris' words, it "looks to a defendant's circumstances, rather than delving the uncertain depths of his character" (2007, p. 527). Specifically, it compares an agent's overt behaviour against that of a hypothetical reasonable person, with normal human capacities, *ceteris paribus*. If we consider Milgram's subjects, it is exceedingly difficult to ascertain their motives, which "apparently manifested tendencies both towards deference to authority and to resist inflicting suffering on others"; given this fragmentation, how are we to "decide that one tendency is more that individual's 'own' than another," especially when "tendencies to act out of character are part of every character" (Doris 2007, p. 526). Experiments in cognitive science compound this problem by calling into question the introspective accessibility of our own motives, which, according to the cited data, tend to be largely confabulated, rationalized, and distorted. Although SRP is not immune from criticism, it has the advantage of relying on objective, testable criteria—namely, whether the defendant had an objective excuse, in terms of either capacities or environment. For example, consider an otherwise upstanding citizen who commits perjury after being threatened by an intimidating Mafia agent. This person can unimpeachably appeal to an excuse of duress. *Duress* is evaluated by considering what a typical human adult is capable of withstanding. As Doris points out, although excuses are not determined entirely by *population base rates*, "still, reflection on base rates helps determine what can fairly be expected of a particular individual in particular circumstances; surely it is partly because most people yield under torture that it seems unfair to hold victims responsible for failing to resist it" (2007, p. 527). Hence the Milgram subjects may have had an excuse, insofar as they all surpassed the expected threshold, and most continued on to the maximum intensity. This at least provides *prima facie* evidence that there were duress-like situational pressures in effect. This applies irrespective of the agents' motives, which, according to Doris, were both obscure and ambivalent. This evaluation departs from 'subjective' theories that only weigh an agent's motives, since base rate statistics are external to the agent. Such theories fail to give these factors any moral/legal valence at all.

One might infer from this explanation that everyone has an excuse, since 'everyone has a story,' as they say. But this would be a mistake. I will discuss this misconception at greater length in section 6, but for present, consider an elucidative example. My position is that a person is excused only if situational factors foreclose alternative deliberative possibilities, or, as Fricker elegantly puts it, they "close off… a more enlightened perspective" (p. 105). Consider two epistemologically different cases, which drive home the moral discrepancy. In the first scenario, a rich business owner declines to donate money to a worthy charity. This person obviously ignored readily-available examples of other business owners who happily donated, such as Bill Gates, the famous corporate philanthropist. Now my claim is that the Milgram participants lacked such counterexamples. At least, none were obviously present. But suppose that the experimenter had informed the subject, on a false pretext, that the previous subject had refused to continue. This

would have provided the subject with another deliberative option. He would now be in an equivalent, or a relevantly similar, position to the rich business owner. If the Milgram subject were to continue on to the maximum threshold under these conditions, he would clearly be blameworthy. Now, it may still be difficult to draw this distinction in many cases, but this is a necessary effect of limited human insight. The same problem beleaguers Williams' account of 'internal reasons,' which holds that a reason is available to an agent as a possible motive for action, if it is accessible through a "sound deliberative route"—an expression that Williams does not explicate (1995, p. 35). This is because this type of question is a matter of empirical, case-by-case assessment. Williams gives the example of a glass of petrol, which the agent mistakenly believes to be water. Since the reason to refrain from drinking the petrol is not available to the agent, we cannot reasonably expect him to be motivated by that reason. So it is likely that he will drink the petrol. But of course, he would not be responsible for drinking it, any more than would a sense-deprived person who cannot detect the poison. In this case, there is *ex hypothesi* no 'sound deliberative route' to a motive to reject the poison, but real-life cases are typically more complex than thought experiments, and resist *ex hypothesi* simplifications. My purpose in mentioning Williams is to show that the problem of assessing cases is not unique to SRP. Is Smith responsible for murder? We will not know the answer until we have a case file containing Smith's personal history, psychological condition, and many other relevant details.

Before moving on, a clarification is in order for metaphysically-minded people. The clarification concerns whether the epistemological problems refers only to an epistemological problem (regarding introspective access to the mind,) or a metaphysical problem (regarding the structure of the mind) as well. The answer, in short, is both. In the first place, 'the epistemological problem' refers to the fragmentation of the mind, which is neither evaluatively integrated nor doxastically coherent. That is, our values and beliefs do not form a coherent, situation-invariant network. In the 1920s, researchers found that character is not consistent from one situation to the next (Hartshorne & May 1928; Newcombe 1929). For example, Harshorne and May found that honest and dishonest behaviours were displayed differently across different situations. This was corroborated by later research, such as Ross & Nisbett (1991). This suggests that, in spite of our subjective sense of integrity, values and beliefs are actually very weakly integrated into a conceptual matrix. Furthermore, the mind's affective processes ('system 1') and computational processes ('system 2') often work at cross-purposes, generating irrational decisions, in order to conserve cognitive energy. This explains our susceptibility to cognitive bias. To illustrate, Kahneman (1982) demonstrates *base rate neglect* by presenting subjects with a stereotypical caricature of a librarian ("Tom W. is of high intelligence, although lacking in true creativity…,") and asking subjects to guess which academic field Tom is likely to enter. Most choose library science, even though the other disciplines are significantly larger (e.g., medicine, social science and social work.) Thus, subjects tend to substitute one attribute, stereotypical resemblance, for the more

significant attribute of baseline percentages. This is a case of what Sinnot-Armstrong et al. call "unconscious attribute substitution" (p. 250). They point out that this type of substitution also affects moral reasoning, wherein we tend to substitute a simpler (more accessible) affect heuristic (e.g., a feeling of wrongness) for a less tractable principle (e.g. the categorical imperative,) resulting in false judgments. Furthermore, on this dual-systems view, there is no central processing unit regulating cognition, which partly explains why the two systems operate at cross-purposes. In a similar vein, Steven Stitch, in 'The Fragmentation of Reason' (1992), argues that people reason in ways that "depart seriously and systematically from what is rational and normatively appropriate" (p. 11). This is evidenced in such cognitive distortions as the selection task (p. 4), the conjunction fallacy (p. 6), and belief perseverence, i.e., the tendency for beliefs to persist when the evidence for those beliefs is no longer accepted (p. 8). Such irrational decision-making, Stitch says, illuminates a systemic rational defect in human cognition.

This should not impugn the idea that in everyday life there is a degree of observable behavioural regularity, however, because in everyday life there is a degree of situational regularity—albeit it a smaller degree than most people presume. Belying accepted wisdom, Kunda and Nisbett (1986) find that laypeople are no better than chance at predicting behaviour across contexts. When subjects were asked to predict whether someone who had displayed honest behaviour and extroverted behviour in two previously scenarios would do so again in a different scenario, they estimated the probably around .81, while the actual probably was only .21. This shows that laypeople's predictions of character are "substantially overconfident" (Doris 1998, p. 523). One may dispute this claim on phenomenological grounds, but it is dangerous to privilege subjective experience above empirical research. As Harman remarks, we also have the intuition that something dropped from an airplane will drop straight down to earth, but we do not allow this intuition to trump the evidential weight of physics. While it is true that some behaviours—perhaps basic evolutionary adaptations—are systematically rational (such as recoiling from a close flame,) many behaviours are systematically irrational, especially those with moral implications. For instance, people are highly susceptible to racial bias, and if one needs tangible proof, this is borne out by the implicit association test (IAT). IAT results indicate that virtually everyone, regardless of one's race, associates white faces with positive words and black faces with negative words (Nosec, Banji & Greenwald 2012). More generally, Kelly et al. (2010) use IAT it to show that, "implicit racist biases can persist even in persons sincerely professing tolerant or even anti-racist views, and… implicit racial evaluations can be insulated in important ways from more explicitly held beliefs" (2010, p. 435). So racial bias, which is a form of irrationality—and, according to Kelly et al., a form of moral dissociation—is both systematic and endemic.

Now, on the basis of this sort of research Doris (2010) infers that behavioural dispositions are both "fragmentary" and "frail" (p. 359); that "cognitive functioning has itself been shown to be

highly susceptible to situational variation" (p. 359); that people typically suffer from "moral dissociation," in which their behaviour defies their moral convictions, which explains the Milgram results (p. 363). Furthermore, Doris rejects attempts at 'unifying explanations,' which attribute rational defects to a single variable, such as fear of embarrassment (viz., Sabini and Silver 2005), insisting instead that, "a great number and variety of inconsequential situational factors" systematically distort rationality (p. 367). Finally, he affirms that, "many important cognitive and motivational processes proceed without intentional direction, and result in cognitive and behavioral outcomes inconsistent with the agent's (often reflectively affirmed) evaluative commitments" (p. 371). In other words, Doris affirms my position that the human mind is deeply (metaphysically) fragmented, and lacks a central control mechanism which would systematize cognitive processes. *However*, Doris does not think that psychological fragmentation is irremediable, or impervious to training. Indeed, it seems that no moral philosopher is this skeptical, given that almost all conclude their work with suggestions for achieving greater internal coherence. This is precisely what Doris does in the 'Moral Psychology Handbook' (MPH). There, he collaborates with Merritt, and endorses her claim from ten years earlier, that we should aim to foster reliably recurring social settings and relationships, as a means of achieving stable behavioural dispositions, i.e., character. So although Doris endorses the metaphysical fragmentation thesis that I have proposed here, he does not take this to refute the existence of narrow traits, or the possibility, to some extent (to be empirical determined,) of enhancing internal coherence. Elsewhere, I have argued that we may be able to achieve dispositional regularity in spite of the systemic irrationality and inscrutability of computational (system-2) processes, by conditioning affective (system-1) processes, through certain types of enculturation, socialization and operant conditioning. This would bypass higher-order cognition and thus avoid problems of cognitive unwieldiness. Then, then even if system-2 processes are hopelessly fragmented and inscrutable, we can nonetheless achieve a degree of behavioural regularity, which, on my externalistic view, amounts to character.

The second interpretation of the epistemological problem is perhaps more self-evident. Namely, it is that we do not have clear epistemic access to our reasoning processes. This should by now be obvious, but the matter is further illuminated by the phenomena of 'moral dumbfounding' and what I shall call 'resistance to introspective access.' Haidt coined the term 'moral dumbfounding' to describe our collective propensity to confidently make moral judgments without being able to provide a general justification for those judgments. For example, when Haidt and Hersh (2001) queried subjects about incest, the subjects insisted that incest was morally wrong, but could not justify their claim. The researchers concluded that the subjects were judging on the basis of intuition, and did not have access to their own reasons or reasoning processes. I have designated 'resistance to introspective access' to refer to the related propensity to deny the common susceptibility to cognitive biases, even when debriefed on them and presented with

extensive evidence of their influencing force. For instance, when Latane and Darley (2007) explained bystander effect to subjects, the subjects still insisted that the number of people present had nothing to do with their behaviour:

> We asked this question every way we know how: subtly, directly, tactfully, bluntly. Always we got the same answer. Subjects persistently claimed that their behavior was not influenced by the other people present. This denial occurred in the face of results showing that the presence of others did inhibit helping. (p. 124)

These examples suggest that our moral reasoning processes are epistemically opaque. As a result, even if we tried consciously to manipulate them, we might not enjoy much success. As aforementioned, I believe that we can enhance cognitive integrity by modulating system-1 processes, but I do not need to argue that here. The aims of the current section are much more modest: to clarify the ontological status of the epistemological problem. The foregoing discussion shows that it is both metaphysical (pertaining to the structure of the mind), and epistemological (pertaining to conscious access to the mind.) Due to the depth and inscrutability of certain types of irrationality, these defects *can* constitute an excuse, but only in the absence of foresight. For instance, if a person is aware of base rate neglect, she is obliged to pay special attention to base rate statistics, and if she fails to be vigilant, she is responsible for the resultant stereotypical judgments. Since stereotyping can have moral implication, this person may be morally blameworthy.

In the next section, I show that SRP is appropriate not only for determining legal responsibility, but also for determining *moral responsibility*. This transposes some of Doris' arguments into the domain of moral philosophy, where, I shall argue, they are equally admissible. In brief, I contend that Holmes' defense of SRP invokes clear-cut moral justifications, which transfer unproblematically into the moral context. These justifications invoke the concepts of social utility and fairness, which must be balanced alongside our considered judgments and theoretical considerations, via the process of wide reflective equilibrium.

**8. Reflective Equilibrium**

To recapitulate for purposes of clarity, Holmes states that "a blind man is not required to see at his peril," yet "when men live in society, a certain average of conduct, a sacrifice of individual peculiarities going beyond a certain point, is necessary for the general welfare" (p. 46). Moran describes this account as one which, "though it does not demand perfection, does insist upon a certain level of prudence or attentiveness to the interests of others" (p. 18). It requires, on the one

hand, that “the accused have both the capacity to required to comply with the law and a fair opportunity to exercise that capacity,” and, on the other, that she “appropriately attend to the interests of others” (Moran, p. 12). This account aims to balance individual rights against the common good, “liberty” against “security” (Moran, p. 6). To illustrate, consider Vaughan v. Menlove (1837), where SRP was first introduced. The defendant built a hayrick on the border of his neighbour’s property and ignored warnings that it was likely to ignite, stating that he ‘would chance it.’ The hayrick predictably caught fire, destroying several neighbouring cottages. On appeal, Menlove argued that he was not responsible, because he simply did not possess “the highest order of intelligence” (Moran 2003, p. 19). Using SRP, the court deemed him criminally negligent. Their reasoning was that the defendant was in fact intellectually normal, and should have known better in light of the warnings. They rejected his defense.

This case elucidates the moral scaffolding of SRP. In broad strokes, the standard attempts to balance individual rights against the general welfare. One can interpret SRP as placing deontic constraints on standard Benthamite-esque utilitarianism (which aims to maximize aggregate happiness,) to achieve an optimal distribution of rights and obligations. The motivating idea is that people have presumptive responsibilities, but it would be not only useless, but also unfair, to submit a person to unreasonable moral demands. One reason for endorsing this model of justification is normative, and the other is descriptive—namely, it presents an accurate descriptive moral psychology for normal human subjects. This is exemplified in research on trolley dilemmas, in which psychologically normal subjects judge that it is permissible to kill one person in order to save five by pulling a signal lever at the intersection of two tracks, diverting the trolley onto a spur (‘spur case,’) but not to accomplish the same result by pushing a large man off a footbridge onto the tracks (‘footbridge case.’) (Cushman et al. 2010). This seems to show that ordinary moral agents make characteristic utilitarian judgments in some moral dilemmas, but characteristic deontological ones in others. In other words, consequentialism is constrained at the limit by deontic reasoning. Interestingly, people who do not invoke these principles tend to be cognitively impaired—for example, researchers observe abnormal moral judgments in people with frontotemperal dementia (FTD,) who exhibit blunted emotions, on the one hand (Mendez et al., 2005), and in people with defects in working memory and abstract reasoning, responsible for controlled cognition, on the other (Greene et al. 2004). The first group cannot generate characteristic Kantian judgments, and the second group cannot generate characteristic consequentialist judgments, resulting in skewed responses. Additionally, Harman, Mason, and Sinnott-Armstrong (2010) argue that on a psychological level, intrapersonal moral reasoning involves “networks of nodes” which receive “positive or negative excitation,” until they settle into a “relatively steady state” (p. 238). This indicates that human cognition most closely resembles reflective equilibrium. Collectively, this research lends credence to a reflective equilibrium model of moral reasoning involving consequentialist and deontic elements. This is not proof of this

model, but it is tentative inductive evidence.

This departs somewhat from the letter of Holmes' writing, but I regard wide reflective equilibrium as the best method for achieving the aims of SRP, i.e., balancing principles (or heuristics) against considered judgments or intuitions and theoretical considerations in an optimal fashion. Thus, it is consistent with Holmes' intentions, and adds methodological specificity. On my view, SRP entails considering utilitarian and deontic principles (which can elaborated in various ways, encompassing considerations of fairness, dignity, personhood, and so on, as per traditional formulations,) intuitions and considered judgments about excusing conditions, and background theory, and adjusting or revising recalcitrant elements to achieve equilibrium. This resembles Rawls' conception, which was introduced partly as a remedy to unacceptable utilitarian conclusions—particularly that one may maximize aggregate utility at the expense of the most socially disenfranchised. Reflective equilibrium will not license such a counterintuitive output, and necessitates adjustment of moral principles (such as utility-maximizing) or theoretical considerations (such as consequentialism.) Similarly, in judging responsibility, it is necessary to balance deontic and consequentialist principles against possible outcomes. For example, if utilitarianism licenses sentencing an innocent defendant to quell public outcry, or blaming someone in psychosis for acting on hallucinatory thoughts, there is *prima facie* reason to adjust the utilitarian principle. This does not mean that we *must* do so, but the counterintuitive conclusion generates pressure to adjust one element or another.

Traditionally, philosophers have defended foundationalist, single-principle moral theories, but pluralist (multiple-principle) accounts are increasingly popular (viz., Goodman 1953, Daniels 1979, Kelly & McGrath 2010). I take Scanlon (2008) to provide the most persuasive defense of value pluralism, beginning with the impossibility of solving moral dilemmas without moral remainder (e.g., ambivalence, regret.) In the classic 'spur' versus 'footbridge' duality, almost everyone believes that it is permissible to pull the lever, but impermissible to push the large person to his death. It is the philosopher's task—perhaps with the help of empirical researchers—to figure out why. Is there an underlying, unifying principle? Perhaps the most common single-principle response appeals to 'the doctrine of double effect' (DDE), whereby it is permissible to cause serious harm as a side-effect of enabling a good consequence, but not as a means of bringing about the same result. That is, DDE proscribes intentionally instrumentalizing people. Yet Scanlon disputes the cogency of this principle, in part because it does not actually motivate our moral judgments (2008, p. 2). Indeed, Cushman et al. (2006) and Waldman & Dietrich (2007) have shown that DDE is not normally invoked in people's explanations of their judgments, and many people are not aware its existence. An additional consideration is that many utilitarians now endorse some version of pluralism, which allows non-utilitarian considerations—typically deontic constraints—to play a corrective role in moral reasoning (e.g., Hooker 2000, Cummiskey 1996, Parfit 2008). This allows utilitarians to justify intuitive responses to classical moral dilemmas.

Scanlon is also concerned that, “there is no single explanation for all of the cases to which the doctrine of double effect is thought to apply. The explanation of what is permissible or impermissible in transplant cases [i.e. another type of moral dilemma] is different from what explains the difference between terror bombing and tactical bombing, and different still from the explanation of what is permissible when dealing with runaway trolleys” (p. 4). Thus, we must invoke several principles to resolve moral problems. The most obvious candidates are consequentialist and deontic reasons, which may encompass diverse considerations—for instance, fairness is typically thought to be deontic, and welfare to be consequentialist. Scanlon endorses welfare and utility as relevant moral reasons, along with other principles that suitably-motivated persons “would not reasonably reject” (1998, p.5). This resembles Rawls’ account in essence, and indeed, when Scanlon explicitly discusses responsibility, he cites avoidability as a central concern (1998, p. 249). We might see avoidability as tied to fairness regarding blame. Indeed, the distinctive feature of moral responsibility, I believe, is that *avoidability* is a central element, which ensures that we do not ascribe blame to people who could not reasonably have done otherwise. This is different from normative ethics, which is not essentially concerned with avoidabilty, because it is not essentially concerned with excuses. For instance, it may be morally impermissible (*tout court*) to torture puppies, but this does not settle the question of whether a particular person is morally responsible for doing so. This latter question depends on a range of factors which attest to whether the act was avoidable.

I will not say much about the metaethical status of wide reflective equilibrium here, as I do not think that I need to take a stance on this question. There are vastly differing interpretations of this method, some realist, some anti-realist, and some lying in between. In spite of pervasive assumptions that reflective equilibrium is a coherentist (non-foundationalist) view, this is not necessarily so. As Noah Lemos (2002) puts it:

> it is important to distinguish the coherence theory of justification from wide reflective equilibrium. The latter, as noted above, seeks coherence between our particular judgments, general principles, and various background beliefs and theories. Still, acceptance of the theory of wide reflective equilibrium does not commit one to accepting a coherence theory of justification. One might value the seeking of coherence without holding that coherence is the only source of justification. So, for example, one might hold that it is reasonable for the scientist to seek explanations that cohere with his perceptual observations and still hold that his perceptual observations have some source of justification other than mere coherence, such as their being grounded in his nondoxastic perceptual experience, or their issuing from a reliable faculty of perception. (p. 500)

Typical candidates for nondoxastic moral experiences in moral reasoning are certain kinds of intuitions, beliefs, or moral perceptions. One influential semi-foundationalist account of reflective equilibrium is Sharon Street's (2006), which holds that evolutionary forces (obviously) inform evaluative beliefs, and that evaluative truth is therefore partly a function of these forces. However, this does not pose a problem for realism, since the realist can hold that although evolutionary forces are prior to moral truth, they factor into reflective equilibrium along with other evaluative judgments, giving rise to rational belief (under ideal epistemic circumstances). Thus, using wide reflective equilibrium, we decide which evolutionary forces are morally relevant. The process is similar to constructing a boat in mid-ocean, by repairing those peripheral planks which diverge from the ideal blueprint (Street, p. 121). Likewise, we adjust peripheral beliefs that diverge from the center. Ultimately, we aim to achieve a coherent framework, constrained by foundational elements.

Now, this weak-foundationalist realist interpretation seems reasonable to me, but I do not need to take sides. For present purposes, I can remain ambivalent about the various metaethical possibilities. What matters is that wide reflective equilibrium, on some metaethical description, is both normatively and descriptively accurate. Thus, it provides a sound methodological basis for implementing SRP on Holmes' description.

That said, I endorse one modification to Rawls' theory, which has been proposed by Kelly and McGrath (2010), viz., to redefine Rawls' conception of 'considered judgments' (which are meant to be weighed alongside principles and background theory,) to be more inclusive. Rawls defines 'considered judgments' as:

> judgments of which one is confident (as opposed to uncertain or hesitant), that are issued when one is able to concentrate without distraction on the question at hand (as opposed to when one is 'upset or frightened') and with respect to which one does not stand to gain or lose depending on how the question is answered. In addition, such judgments must be stable over time. (Kelly and McGrath, p. 335).

The problem is that this definition rules out many judgments that are patently valuable and informative, especially from the perspective of social and feminist epistemology—disciplines concerned to include historically-marginalized testimony in philosophical discourse. Kelly and McGrath assert that although *impartial* testimony may appear *prima facie* to comprise 'ideal epistemic conditions,' *partial* testimony is often epistemically valuable: "Not all unreasonableness is due to the operation of the kind of general corrupting factors (e.g., being personally invested in how a given question is answered) that the relevant conditions exclude" (p. 335); rather, a personal stance can contribute valuable information to the resolution of a moral question. The authors emphasize that for a racial minority in the original position, race may be an important consideration, and so it should not necessarily or automatically be excluded from deliberation. A

person in this position may in fact have credible evidence for saying that, say, there should be affirmative action policies, which inherently take race into consideration. This is only part of Kelly and McGrath's critique, but I take it to be their key contribution, and certainly the most relevant for present purposes. Accordingly, in order to construe wide reflective equilibrium inclusively, we must allow *partial* testimony to play a role in moral reasoning. To give a specific example from personal experience, the Empowerment Council is an independent organization consisting of psychiatric clients who inform policy at the Centre of Addiction and Mental Health (CAMH,) including ensuring the representation of client perspectives in CAMH policies, educating, informing, and sensitizing mental health professionals, conducting outreach and community development, and various other functions (http://www.empowermentcouncil.ca.) Clients have a vested interest in mental health policy, and yet this should not preclude them from deciding on principles of justice regarding mental health systems. Rather, this partial perspective—what we might call 'lived experience'—renders their knowledge, disseminated through testimony, particularly epistemically worthy. So while impartiality may comprise an ideal (hypothetical) epistemic standard, it does not reflect a pragmatic epistemic standard, which we ought to adopt in real-life reasoning. Rather, partial testimony should be included, and should furthermore be actively elicited on the part of social decision-makers and authority figures. Note this this view fits with Fricker's account of (partial) testimony as valid evidence for moral belief. What this means for moral responsibility is that when contemplating how to deploy the reactive attitudes, we should take into consideration those who have a personal stake in certain blaming practices. Indeed, we may even give partial testimony particularly high credence, since it involves lived experience.

This methodological discussion should suffice in context of the present inquiry. The next matter of concern is the role of deontic and consequentialist principles in reflective equilibrium. The first thing to note is that the fairness condition—which is perhaps the most controversial element—is required for action-guidingness, since ignoring the agent's ability to avoid negative consequences—that is, ignoring the agent's psychological capacities and environmental affordances—would generate a motivationally inert standard. There is ample support for this requirement, particularly in scholarship on 'reasons internalism' (which, in spite of the title, is consistent with moral-responsibility externalism, as I have construed it here). Reasons internalism is defined by Williams as the view that moral claims should connect with the agent's "subjective motivational set" through a "sound deliberative route" (p. 35). The sound deliberative route, of course, must be defined partly in terms of *external* factors, such as whether a relevant reason was available in the agent's epistemic environment, which is something that Williams conveniently elides, leading to significant confusion and disagreement. My account of responsibility *includes* the condition that a responsible agent must have normal motivational capacities (which are observable,) but also comprises situational factors. According to SRP, if these conditions are not met, it would be unfair to deploy the negative reactive attitudes, thereby

inflicting social harms. Another way of putting this is to say that ascribing blame in the absence of the ability to do otherwise would amount to browbeating, i.e., inflicting useless punishment. This action objectifies the agent by treating her as a mere means to the satisfaction of one's punitive impulse. Moreover, there can be no utilitarian justification of this practice, since the sanctions make no sense to the agent, and cannot elicit an urge to make amends, pay reparations, or ask forgiveness.

There is additional support for this principle in experimental philosophy, where the majority of scholars have (implicitly) signed on to some version of Owen Flanagan's (1991) principle of minimal psychological realism (PMPR,) according to which a viable ethical theory must "draw on an image of ourselves that we are capable of admiring and to which we can in some sense imagine conforming. A normative conception which fails to meet certain standards of psychological realizability will fail to grip us, and in failing to grip us will fail to gain our attention, respect, and effort" (p. 26). This principle is reflected in the work of the Moral Psychology Research Group (2010). For example, Doris says that central questions in the field of ethics "want empirically informed answers," which only scientific researchers are in a position to provide (p. 1); Prinz and Nichols state that, "empirical moral psychology is essential for moving from imprecise formulations into more detailed explanations of [the nature of moral emotions]" (p. 111); And Sinott-Armstrong, Young, and Cushman lament that, "until the last decade of the twentieth century, philosophers and psychologists usually engaged in their enterprises separately," preventing them from determining "whether certain moral intuitions are justified"(p. 247). These philosophers agree that a viable moral theory must be motivationally efficacious, else it cannot play a role in our moral practices. If so, then my proposal is on the right track. The fairness condition embedded in the proposed account of wide reflective equilibrium constrains responsibility ascription to cases of potential response, and hence respects PMPR.

Michael Smith gives us reason to worry that motivationally-inert moral claims may not only fail to motivate, but can cause significant social harms. (Thus, deontic and consequentialist reasons are intimately reticulated.) Following Gary Watson's example (1975), Smith asks us to imagine that we have just been defeated in a game of squash, and we are so humiliated that we are consumed with a desire to smash our opponent in the face with our racquet. If we were fully rational, we would stride across the court and shake our opponent's hand in a show of good sportsmanship, but in our current state, we would not be able to resist violently assailing the person. It follows, says Smith, that what we have utmost reason to do in the circumstances is "smile politely and leave the scene as soon as possible" (p. 111). This is not what our fully rational self would do, but it is what our fully rational self would *advise* us to do, in light of our imperfect motivational profile. This constitutes what Smith calls the 'advice model' of moral motivation. The same argument, I believe, shows that moral responsibility ascriptions should be sensitive to a person's motivational capacities. For example, if we are familiar with a person's

anger-management issues, and suspect that blaming him for failing to show good sportsmanship would only goad him into trying, in vain, to shake his opponent's hand, and inadvertently smashing the poor victim in the face against his better judgment, then this is not a reasonable blaming policy. What is more reasonable, perhaps, is suggesting psychological counseling, recommending a less invigorating hobby, or, failing this, warning the opponent of the man's psychological disorder. Similarly, if imposing an immediate 10% corporate income tax—plausibly a form of reactive attitude—would inadvertently motivate executives to move their operations offshore, or incite a libertarian reaction against the whole system of progressive taxation, then this is not a reasonable course of action. However, if imposing a 3% tax increase for ten years would motivate most people to comply, it would be advisable. In this way, SRP recommends a pragmatic approach, which takes into account motivational capacities, and the effect of blaming practices on general wellbeing. Notably, although motives are psychological states of the agent, they constitute part of what Smith calls "circumstances," as in the statement: "the desirability of an agent's Φing in certain circumstances C depends on whether she would desire that she Φs in C if she were fully rational" (p. 110). C is equivocal between psychological and situational elements, and most philosophers neglect the possible extenuating force of the latter. Thus, it is important to specify that C includes situational factors, such as whether the agent had been encouraged to seek psychological counseling in the past. If so, then he deserves blame after all. On the other hand, if he lives in a country where psychiatry services are not covered by social insurance, and he cannot reasonably afford the cost, then he may not be blameworthy, since his government has failed him. This might mean that an American and a Canadian with anger-management problems bear different degrees of responsibility in spite of their psychological similarities, based on their nationality alone.

Another reason for thinking that motivationally inefficacious blaming is unfair is that it places unfair burdens on the morally least well-off. If we were to step behind a Rawlsian 'veil of ignorance' and decide on just principles, not for structuring society, but for *deploying the reactive attitudes*, there is reason to think that we would withhold negative reactive attitudes from those who lack minimal moral competence due to an unjust distributive scheme. Of course, we ought to distribute moral resources fairly in the first place, but failing this, it may be justified to withhold blame from those who have been egregiously morally deprived. To see this, we can sketch a rough genealogy of moral development. To begin, it seems that certain goods are required to develop minimal moral competence, and if these goods are absent at a critical phase of human development, there is a commensurate likelihood that an agent will become morally incapable—perhaps psychopathic. This premise is supported by figures from Statistics Canada showing that, "individuals from households with low incomes were more likely than those from high income households to report socially disruptive conditions in their neighbourhoods," and these conditions

"may reflect crime levels in their neighbourhoods" (http://www.statcan.gc.ca/pub/85f0033m /2009020/findings-resultats/f-r5-eng.htm). While this only proves correlation, it inductively suggests that there could be a causal relationship between poverty and criminal behaviour, which may in turn be excused by the situation. (Indeed, on a situationist picture, it is certainly more likely to be excused than on a character-based account, which would methodologically neglect the impact of broader systemic factors.) If the posited causal relationship holds, *and* this relationship is due to an unfair distribution of relevant resources (material necessities, moral education, positive role models,) then it seems reasonably to say that those who failed to distribute the requisite resources fairly are partly responsible for the outcome, and the victims of the unfair distributive scheme are commensurably excused. In effect, this is a situation of coercion, in which situational pressures, partly created by the most well-off, exert exculpatory pressures on the least well-off. On this picture, there is an inescapable responsibility on the part of authority figures—especially those who control morally-relevant resources such as basic necessities, social programs, early childhood education, etc.,—to distribute such resources equally, to the greatest extent possible, so that the fewest number of potential moral agents fall outside of the ambit of the reactive attitudes (to whatever extent.) This point is made, albeit indirectly, by Nussbaum (1993), who developed a now-famous 'capabilities approach' to human functioning and flourishing, which posits ten basic human capabilities that governments are morally required to foster in all people (regardless of nationality.) These include good health, adequate nourishment, adequate shelter, being able to use the five senses, being able to form a conception of the good life, and so on. These capabilities are required for basic human functioning, and, in greater quantity, human flourishing. This is a supplement to the Rawlsian materialist account of distributive justice, which, Nussbaum says, neglects important distributive resources—moral resources that are required for a minimally human life. Extrapolating from this account, I infer that until such a time as *moral* resources are equitably distributed, it is unfair to blame the morally least well-off. Furthermore, the most well-off are obliged to distribute these resources, and insofar as they do not, they are partly responsible for the result, i.e. pervasive moral incompetence. This is extends Rawls' distributive theory, applying the distribution principle to *moral* resources, and entailing greater burdens of responsibility for the most well-off. This view, furthermore, explains our sense of *collective* responsibility, which emerges from an acknowledgement of the importance of interpersonal relationships for generating obligations. And it handily justifies the intuitive claim from the Spider-Man comics, that 'with great power comes great responsibility.'

Unlike agent-centric theories, SRP gets the distribution of responsibility right. Consider battery. On my view, a person is personally morally responsible for battery, *unless* the action was preempted by situational factors. This may be the case in some prisons, where violent crime is substantially higher than in the general population. Although I could not find comparative statistical data for prison inmates versus non-incarcerated Canadians, media reports indicate that

violent assault in Canada's prison system affected 1669/14,400 inmates, or 8.09%, in 2011/12,[3] while Statistics Canada finds that violent crime affected 1190/100,000 Canadians, or 0.01%, in 2012.[4] This is a significant discrepancy in exposure to violence. In addition, prison violence is at a "historic high," according to prisoner ombudsman Howard Sapers, due to overcrowding. CBC News reports that, "in the past two years, 1000 new inmates entered the system, even though there were no new beds… The influx has led to an increase in double bunking, even in maximum-security units… In [some] cases, Sapers said, some inmates have been forced to share a cell that is less than five square metres…. Sapers told CBC News that overcrowding has led to growing tensions and violence" (http://www.cbc.ca/news/canada/record-high-prison-numbers-sparking-violence-1.1260764). If so, then prisoners may effectively be in a situation of duress, which may mitigate culpability for violence. According to externalism, the architecture of prison systems could constitute a mitigating factor in cases of assault, as it did in the Stanford Prison Experiment. Externalism does not neglect psychological capacity (based on observable evidence,) but it emphasizes situational factors as potential excusing conditions, regardless of the agent's psychological profile. For example, if the prison system breeds prisoners who enjoy harming others, and this subjective enjoyment was provoked and reinforced by years of incarceration under strict, top-down codes of conduct and police brutality, then the agent is excused *regardless* of his character, choices, or intentions. This explanation applies to Zimbardo's participation in the defense of Staff Sergent Ivan Frederick, who served as a guard in Abu Ghraib Prison, and was accused of multiple crimes, such as maltreatment of detainees, assault consummated by battery, indecent acts, and dereliction of duty. Zimbardo's rationale was that Frederick deserved a mitigated sentence in light of the prison conditions. He highlights several coercive factors, some of which echoed SPE, such as nonstop night work and exhaustion; constant fear of retaliation from prisoners, who were often armed and in collusion with Iraqi guards; lack of mission-specific training; lack of oversight; lack of contact with the outside world; initiation into "extreme methods of interrogation and secrecy" (p. 345); and dungeon-like living conditions. Zimbardo emphasizes that prior to entering Abu Ghraib, Frederick was an outstanding American citizen. All evidence supports the hypothesis that he was corrupted in Abu Ghraib, and transformed into a moral monster. While he served in Abu Ghraib, there is no question that he was characterologically heinous, or that he made flagrantly impermissible decisions, but his actions deserve mitigation in light of purely external forces. Zimbardo concludes, and I agree, that the Bush administration bears the brunt of the responsibility for the atrocities that occurred in Abu Ghraib. What Zimbardo lacks is a philosophical defense of this judgment, which I have provided here. Zambardo's intuitions are correct, and accord with moral-SRP.

---

[3] http://www.thestar.com/news/canada/2012/07/07/star_exclusive_violent_assaults_in_federal_prisons_on_ the_rise.html

[4] http://www.statcan.gc.ca/pub/85-002-x/2013001/article/11854-eng.htm?fpv=269303#a1

How does this standard compare to character-based accounts of responsibility? Consider one very influential example: Frankfurt's identificationist account, which is bizarrely endorsed by Doris. On this view, an agent is responsible for actions that she endorses at the level of her higher-order volitions (or deep self)—that is, actions that she wholeheartedly endorses. Did Frederick wholeheartedly endorse his heinous actions? Yes. Is he responsible? If Zimbardo is correct, then no. Why not? Because he acted under situational coercion. He went from being an upstanding American citizen to a moral monster in a matter of weeks (if not days,) and it is reasonable to suppose that the Bush administration, who ratified the prison's policies, are largely responsible for decimating Frederick's moral fiber and compelling him to perpetrate crimes. Their role in the Abu Ghrahib atrocities was similar to that of the proverbial Mafia intimidator: Frederick was in constant danger of physical battery, constantly sleep deprived, and extremely isolated. He was partly a victim of administrative incompetence and gross neglect, for which the administration are culpable. Frankfurt's account has absolutely no means of accounting for this distribution of responsibility. In fact, it gets things absolutely backwards: since Frederick was monstrous, and the prison administrators were negligent, Frederick's psychological profile suggests that he should be more blameworthy. But when we trace the agent's causal history, we come to a different conclusion. The administrators are far more blameworthy.

Another issue with views such as Frankfurt's, which define moral responsibility in terms of the agent's psychological hierarchy (i.e., correspondence amongst the agent's first-order and higher-order volitions,) is that they generate radically counterintuitive excuses, which suggests that they cannot be right. For instance, suppose that you are a teacher, and your student approaches you after a deadline, saying that she spent the whole weekend playing video games instead of doing her assignment, but she does not endorse her actions at the level of her higher-order volitions, and wishes that she had done the assignment instead. Does this have the makings of an excuse? Obviously not. But this cognitive dissonance is, for Frankfurt, an extenuating circumstance! Strawson, by contrast, offers a list of intuitive excusing conditions, which, however, are not obviously commensurable on the basis of a single rationale. However, we can explain all of them using explanatory heuristics such as SRP, which comprise a set of reasonable principles and judgments. This explanatory framework balances moral principles and intuitive assumptions in an optimal way.

Now let us consider the alternatives to SRP in more general terms, based on Wolf's helpful diagnosis. The main alternative—indeed, the standard view in moral philosophy—is what Susan Wolf calls the 'deep-self view,' which is analogous to character theory in legal philosophy (though Wolf does not draw this comparison.) This view is attributable, in its most recognized formulations, to Frankfurt, Watson, and Taylor:

> For Frankfurt, [being responsible] means that our wills must be ruled by our second-order desires; for Watson, that our wills must be governable by our system of values; for Taylor, that our wills must issue from selves which are subject to self-assessment and redefinition in terms of a vocabulary of worth (Wolf 1987, p. 375).

These theorists "agree that if we are responsible agents, it is not just because our actions are within the control of our wills, but because, in addition, our wills are not just psychological states in us, but expressions of characters that come from us, or that at any rate are acknowledged and affirmed by us" (Wolf 1987, p. 375). In this way, character plays a central role in this theoretical purview. Wolf shows that this approach is problematic not only because it is susceptible to the epistemological problem outlined above, but also because it rules out intuitive excuses which are rooted in our "pretheoretical intuitions" (p. 382). Specifically, although it allows us to differentiate between responsible agents and victims of hypnosis, brainwashing, and so on, it does not allow us to excuse deprived childhood victims who are intuitively non-responsible (to some extent.) Wolf illustrates this point with a now-famous example involving "JoJo, the favourite son of Jo the first, an evil sadistic dictator of a small undeveloped country" (1987, p. 379). JoJo follows in his father's footsteps and goes on to torture and imprison people on a whim. Although JoJo has "the desires he wholly wants to have" (p. 379)—which renders him responsible on the deep-self view—he is not responsible because he is not "sane," i.e., he lacks "the ability cognitively and normatively to understand and appreciate the world for what it is" (p. 387). He is "unavoidably mistaken" (p. 383). Although Wolf does not explicitly invoke SRP, her description can be interpreted in terms of the avoidability principle, which states that a responsible person must be able to comply with relevant norms. Since JoJo *ex hypothesi* cannot comply in virtue of his environment, he is not fully responsible. (This is not to say that he should not be detained for the safety of all, but this utilitarian judgment does not entail responsibility. In current legal practice, JoJo would likely be sentenced as not-criminally-responsible under the mental disorders defense [or the equivalent outside of Canada,] and detained in a forensic psychiatry unit.)

This is a fairly common excusing condition in philosophy, which has been cited, under various linguistic guises, by Strawson (1963), McKenna (1998), Wallace (1998), and Fricker (2007). I will use Fricker to illustrate the force of this excuse. Her explanation is useful because it connects this intuition to a deeper moral intuition: the familiar notion that *ought-implies-can*. Fricker gives an example similar to Wolf's. She cites Herbert Greenleaf from Anthony Minghella's *The Talented Mr. Ripley,* whom she sees as partly excusable. In the play, Greenleaf (incorrectly) distrusts the testimony of his son's girlfriend Marge, although Marge's intuitions are accurate. Greenleaf is influenced by the pervasive sexist prejudice of his culture and childhood circumstances, and consequently lacks the requisite "gender-critical insights" to appreciate Marge's truthful testimony (2007, p. 100). Although Greenleaf is sexist, he is not (fully) morally

culpable, says Fricker, because he is not "in a position to know better" (p. 100). His ignorance is due to a gap in the 'shared hermeneutical resources' of his cultural-historical context, in Fricker's words. Thus, Greenleaf represents "a special case of 'ought' implies 'can,' since in our example the 'can' part is a matter of whether Greenleaf could reasonably be expected to achieve the critical perspective on gender that would have enabled him to question his lack of trust in Marge in the requisite way" (p. 101). Under the circumstances, he cannot achieve this level of insight. This analysis echoes Holmes' argument: because Greenleaf does not have access to critical hermeneutical resources, he (epistemically) could not have done otherwise. That is, his actions toward Marge were unavoidable, and he is not fully morally responsible. The epistemic reading of avoidability is contentious, but it is not farfetched. If intellectual disability is constraining, then epistemological should be constraining, provided that *avoidability* is the central evaluative criterion.

If one does not accept this epistemological reading of ought-implies-can, one can still accept a stricter application of SRP, as excusing of physical and intellectual deficits. This, at least, is a fairly uncontroversial interpretation. Admittedly, intuitions diverge regarding *moral* incapacity. However, I believe that the burden of proof is on the person who wants to draw a distinction between physical and intellectual constraints on the one hand, and epistemic constraints on the other, which are not patently disanalogous. If the evaluative criterion for non-responsibility is avoidabilty, then an *ad hoc* reason is needed to justify a distinction in kind (as opposed to degree) between incapacities. Character is the most likely candidate, but I have disputed this evaluative criterion, and will raise further objections in section 8. It may also be worth noting that many philosophers implicitly endorse some form of epistemological constraint on responsibility, such as Nelkin, who construes Kant's principle of ought-implies-can such that moral responsibility requires 'action-directedness,' which is 'built into the very idea of obligation' (p. 114). Thus, on Nelkin's view, those who lack the "perceptual, cognitive, and emotional" capacities required to respond to moral reasons—i.e. psychopaths—are not morally responsible (p. 76). But again, if one denies this reading, one can still endorse a weaker version of SRP, which, on some ground or other, excludes epistemic constraints.

**9. A Caveat: Emotional Awareness**

Before proceeding, I would be remiss if I did not consider a possible lacuna in Fricker's account, which was pointed out to me by Carla Bagnoli in a presentation and subsequent lengthy discussion at the Moral Epistemology summer course at Central European University (CEU,) Budapest (2014).[5] Bagnoli, as I interpret her, agues that testimonial and hermeneutical injustice

[5] All quotations in this paragraph refer to Carla Bagnoli's presentation slides from the Moral Epistemology course (2014). They also reflect her lectures and forthcoming publications,

do not necessarily exclude moral responsibility, due to the possibility of "emotional awareness," which typically "enters practical reflection by [a feeling of] emotional distress." Emotional awareness may take the form of "residual prejudices," which are "inconsistent with [an agent's] actual beliefs." These are the "most interesting cases" of prejudice, because they constitute "the most surreptitious and psychologically subtle forms of testimonial injustice." Bagnoli believes that we can counteract residual prejudices by highlighting their dissonance with other mental states, and so dissonance is "a crucial ethical and epistemic resource." If this is right, then, *contra* Fricker, a person can be morally responsible for wrongdoing even in a climate of hermeneutical injustice—and *a fortiori* in a climate of testimonial injustice—if the person possess "emotional awareness" of pervasive prejudice, and fails to counteract her own residual prejudices to the best of her ability. For example, in Kazo Ishiguro's *Remains of the Day* (1989), the butler, Stevens, dismisses two Jewish girls from his staff at the behest of his Nazi employer. It is unclear whether he *knew* that dismissing the girls was wrong, but assuming that he did not, the decision clearly did not sit well with him, and he expressed ambivalence. This suggests that even if he did not possess moral knowledge, he possessed *emotional awareness*. Accordingly, on Bagnoli's view, he may be partially responsible for his action. Nonetheless, the fact that he did not possess a *belief* that he was doing wrong may constitute at least a partial excuse, since it is difficult to act on non-propositional 'awareness' in opposition to one's explicit beliefs. However, I agree with Bagnoli that emotional awareness imputes some degree of culpability. Thus, if someone possesses the affective resources required to respond to injustice and fails to employ them, the person is *pro tanto* blameworthy.

This does not negate what has been said so far, but it necessitates a more nuanced interpretation of avoidability, and closer scrutiny of particular cases. Specifically, to determine responsibility, we must now attend not only to the person's hermeneutical environment, but also to the person's emotional resources. Normally the two go together, but it is conceivable that they can come apart, especially in situations of moral transformation or learning, when metacognitive (affective) states may precede full-fledged knowledge. In such situations, a person may bear responsibility for failing to utilize emotional resources to the greatest extent possible. That said, there is an affinity between *hermeneutical* resources and *emotional* resources, insofar as circumstances that obscure moral belief may also obscure emotional sensitivity. For example, physical abuse might inculcate not only false moral beliefs (about the acceptability of abuse,) but also emotional deficits or desensitization. Furthermore, statistical data on the correlation between physical abuse in childhood and subsequent aggression may constitute an intrinsic excuse, irrespective of the agent's psychology. Consider, for example, Salzinger, Feldman, Muriel, and Rosario's (1993) report that, "recent reviews concur that physically abused children's behaviour is

---

'Epistemic Oppression via Moral Judgment,' Lecture at Turku University, unpublished ms., and 'Gaslighting: a case of epistemic injustice,' Keynote address at University of Pavia, unpublished ms. I am grateful to her for referring me to these resources.

more disturbed than that of non-abused peers" (p. 169):

> the behaviour effect best documented by both direct observation and parent and teacher ratings is that abused children are more aggressive than nonabused children (Bousha & Twentyman, 1984; Burgess & Conger, 1978; Egeland & Sroufe, 1981; George & Main, 1979; Hoffman-Plotkin & Twentyman, 1984; Howes & Espinosa, 1985; Kaufman & Cicchetti, 1989; Kent, 1976; Kravic,1987; Lahey,Conger,Atkeson, & Treiber, 1984; Reid, Kavanagh,& Baldwin, 1987; Reid, Taplan, & Lorber, 1981; Reidy, 1977; Salzinger, Feldman, Hammer, & Rosario, 1991; Wolfe & Mosk, 1983)… Insofar as development is hierarchically structured, with later functioning depending to a significant degree on earlier developmental achievements, abuse at a given age must be considered not only for its immediate effects on the child's functioning but also for how the changed behavior influences the child's reactions to new situations. It is thus particularly important to address the problem of ag- gressive behavior in abused school-age children, because studies of delinquents (Lewis, Mallouh, & Webb, 1989; Loeber, Weissman, & Reid, 1983; Tarter, Hegedus, Winsten, & Alterman, 1984) and of adults abused as children (McCord, 1983) suggest a robust association between early abuse and continuing aggressive and antisocial behavior (Lewis et al., 1989). (p. 170).

The correlation between childhood abuse and adult aggression may provide evidence of an excuse, just as human susceptibility to waterboarding indicates an excuse for yielding information under waterboarding torture. Thus, while Bagnoli's point is relevant, there may be situational excuses that are independent of psychological considerations. This is why a metaphysically wide construal of excuses such as moral-SRP is needed.

In the next section, I attend to the most likely general objections to this account.

**10. Objections**

As I have emphasized, very few philosophers, including situationists, have addressed the implications of social psychology for moral responsibility, and nearly all have focused on the question of settled character. One notable exception, however, is Miller et al. (2002). These authors identify two principal objections to the situationist approach: (i) that social psychological explanations condone obedience to malevolent authority, which carries a "clear potential to cause harm" (p. 312), and (ii) that social psychology fails to distinguish between those who authorized certain malevolent actions (such as genocide) and those who merely followed orders. Miller et al. try to defuse these worries by contending that, "by exonerating perpetrators in terms of advancing

a situational causal attribution, social psychological explanations in fact convey a more skeptical and excusatory posture toward the propensity of most persons—that is, ourselves—for committing harm" (p. 331). Thus, situationism prompts us to take responsibility for our personal role in wrongdoing, no matter how minimal or remote. I do not dispute this conclusion, but Miller et al. do not explain very clearly how situationist explanations accomplish this perspectival shift, and, perhaps more importantly, they do not offer a direct response to objections (i) and (ii). They merely note that these objections are a source of legitimate controversy, and go on to contend that familiarity with situationism will, by some mechanism or other, enhance our sense of personal responsibility. Or so one hopes.

Let us address (i) and (ii) in turn. Miller et al. cite David Mandel (1998) as the primary protagonist of (i). Mandel says:

> The many oversimplified statements about the Holocaust that have been made by Milgram and a number of other social scientists, like the claims of most accused Nazis, constitute little more than an obedience alibi. The term alibi is especially fitting because it connotes both an excuse or assurance of innocence and an explanation or statement. Holocaust perpetrators have asserted the obedience alibi as an assurance of their innocence. Social scientists have asserted the obedience alibi as an ostensibly situationist explanation of the Holocaust. Though the intent of one group has differed from the other, the message conveyed has been strikingly similar (Miller et al. 2002, pp. 311–312, citing Mandel 1998, p. 91).

Mandel goes on to say that the obedience alibi is offensive to victims and likely "to cause harm" (p. 312). While I do not dispute that social psychology explanations may be perceived by some as a type of excuse, I believe that philosophers exaggerate the extent to which they warrant such an interpretation. The examples above—of JoJo and Herbert Greenleaf—present extremely rare cases of circumstantial epistemic bad luck. JoJo, first of all, was raised by an evil and sadistic dictator in a small, undeveloped country. *Ex hypothesi*, he did not have *any* reasonable opportunity, throughout the duration of his personal development, to acquire moral agency. (At least, this is what Wolf's example suggests.) The vast majority of people are not nearly this culturally insulated. Herbert Greenleaf was much less malicious than JoJo—he merely failed to believe Marge's credible testimony—and according to Fricker's example, he is excusable only because he grew up in a climate saturated with sexism, in which enlightened gender perspectives were historically unavailable. Notably, this type of excuses cannot be applied to individuals living in mainstream Western society. It would be a misconstrual of Wolf's and Fricker's accounts, and of SRP properly understood, to excuse the majority of people, or even a significant number of ordinary citizens, from responsibility in a typical modern democracy. On my view, as outlined in

the previous six sections, we are to judge responsibility by measuring an agent's behaviour against that of a reasonable person, constrained by population base rates and relevant biographic details (e.g., blindness, intellectual disability, and so on.) This standard would potentially excuse social outliers, but certainly not most people. It may excuse, say, a child soldier such as Ishmael Beah (whom I will discuss shortly,) but not a corrupt business person such as Richard Syron, the CEO of Mortgage loan corporation Freddie Mac during the subprime mortgage crisis, which cost many people their homes. On Francesca Gino's (2013) account, Syron ignored the advice of multiple advisors on multiple occasions, including that of the company's chief risk officer, David A. Andrukonis, who warned that many of the loans the company had bought "would likely pose an enormous financial and reputational risk to the company and the country" (Gino, p. 20). But according to Andrukonis, Syron turned a blind eye, and insisted that "we couldn't afford to say no to anyone" (Gino, p. 20). It is difficult to see how 'unavoidabilty' would apply in this case, where Syron's wrongful decisions flew in the face of ample counterevidence and warning signs. Gino's analysis and blaming stance seem eminently reasonable.

I would also point out that Mandel tends to equivocate on the term 'innocent,' employing it in two different senses: (i) to mean that a person is not responsible, and (ii) to mean that a person is not morally flawed. I agree with the second usage, but not the first. I do not think that Hitler was morally innocent, in the characterological sense of the word, assuming that characterological language makes sense. I have meticulously avoided addressing the subject of settled character, but even situationists such as Doris affirm that there is situation-sensitive (narrow) character, and thus a person in recurrent circumstances may be recurrently 'evil,' measured in terms of the chronic instigation of malevolent harm. Since Hitler's life is now a *fait accompli*, we can say, in the situationist sense, that he was overall an evil person, measured by the sum of his actions. What I disagree with is the idea that everyone who perpetrates harms is responsible for doing so. There may have been German civilians who were in some sense complicit in pernicious activities, but who could not have done otherwise in any practical sense of the word. Whether these people were 'bad' is a separate question, but their responsibility depends on the availability of opportunities for defection. Similarly, most of the Milgram subjects behaved badly by their own lights, but this does not necessarily entail that they were acting *responsibly*. It is both sensible and useful to separate these two notions.

This point has important implications for Mandel's worry that social psychological explanations may cause social harms. My view is that the extenuating implications of situationist psychology have been greatly exaggerated. Social psychological explanations, on my understanding, do not excuse most people. They may excuse passive bystanders in certain types of coercive (broadly construed) situations, depending on the particularities of the case. But even if they do, this does not bar us from saying that complicity in antisocial behaviour is objectively

wrong and should be avoided, discouraged, and counteracted whenever possible. The fact that individuals may not be responsible for committing a harmful action does not negate the deontic (wrongness) status of the action, but it shifts the etiological analysis onto sociological factors, which are effective causes. It should be noted that I have nowhere asserted that Nazi behaviour was either good or justified to any extent whatsoever, but I believe that it is important to draw a distinction between *permissions* and *excuses*: to say that an action is excused is not to say that it is morally permissible, or *vice versa*. Excuses acknowledge wrongdoing but identify mitigating factors. On reflection, I believe, we are forced to acknowledge this conceptual distinction, even if we confuse our terminology in ordinary conversation, and are imprecise even in scholarly discussions. It may be this conceptual unclarity that prompts philosophers to draw excessive, exaggerated conclusions from the situationist literature. Indeed, Miller et al., although they support situationism, nonetheless tend to equivocate between the language of 'condoning' and 'excusing.' For instance, they say that they "would expect a relatively condoning image of the perpetrator contained in social psychological explanations of harm or misconduct" (p. 307); that Zimbardo's view "implies a relatively exonerating or condoning stance" (p. 314)—which Zimbardo would adamantly deny; and that "one sees the clearly unpopular effect of the condoning implications of a social psychological perspective" (p. 314). They also conflate the terms 'forgiving,' 'exonerating,' and 'absolving' throughout their book. It is precisely this type of conflation that provokes the hasty conclusion that situationism implies pernicious social effects. Specifically, this terminological muddle suggests that judgments of moral responsibility and judgments of moral permissibility are equally sensitive to ought-implies-can considerations, when in fact only the former are. Hinging permissibility on such a principle would considerably deplete normative ethics, and unseat commonsense morality. However, if we respect the distinction between responsibility and permissibility, then we can regard situationism as modestly revisionary concerning moral responsibility, but not revisionary at all concerning normative ethics. In this case, it would not be as deflationary as critics presume. Moreover, this reading of situationism provides an efficient solution to systemic moral infractions, which is to identify and rectify systemic causal factors. This is, in fact, the implication touted by Doris (1998), Merritt (2000), Gino (2013), and Kahneman (2012), amongst others.

Regarding (ii), Miller et al. cite Leonard Berkowitz (1999) as its main proponent. Berkowitz presents a "highly critical analysis" (Miller et al., p. 313), which accuses situationism of failing to distinguish between those who authorized the Holocaust and those who merely followed their orders. Berkowitz objects that,

> Social psychology's relative inattention to the great atrocities committed during the extermination program reflects the field's failure to establish a conception of evil that differentiates among categories of wrongdoing. In so doing, there is a danger of trivializing

terrible actions (p. 250, quoted in Miller et al., p. 313).

This statement, again, rests on a conflation of normative terms. Situationists do not believe that the Holocaust is on par with telling a white lie, and I certainly have not asserted anything to this effect. I have not contested the normative framework that ranks atrocities as worse than minor transgressions. In fact, insofar as SRP aims to balance utilitarian concerns against individual rights, it requires us to consider the magnitude of the harm (i.e., the disutility.) Thus, a ranking of aggregate disutility is entailed by the view, though it must be contrasted against the agent's capability to avoid or prevent a given disutility. Far from endorsing a nihilistic picture, I have offered a positive and substantive account of moral responsibility, which preserved commonsense intuitions about collective rights and obligations. This view does not threaten moral philosophy, but rather enhances it by clarifying common confusions. Moreover, it is simply false to say that situationists have not paid sufficient attention to the great atrocities. The most famous situationist experiment of all time—the Milgram Obedience Experiment—was specifically designed to investigate what causal factors brought about the Holocaust, and whether they could be replicated elsewhere. Subsequent situationists (previously mentioned) have developed ethical programs that use this research to help us improve our normative capacities. Thus, contrary to Berkowitz's objection, it is more accurate to say that atrocities have been the *guiding concern* of situationist research, rather than a neglected locus.

**11. Clarifications: SRP Versus Internalism**

In this section, I shall make some clarificatory remarks about the difference between SRP and character-based theories. To this end, I will foreshadow arguments which will be expanded upon in later chapters. My aim here is to articulate aspects of these arguments which are useful for demarcating the boundary between SRP and character theories. It should be noted at the outset that SRP is not committed to dispensing entirely with psychological states, as these may be relevant to evaluating avoidability—the core justificatory principle of SRP, on my conception. Hence, insofar as a given psychological deficit preempts behaviour unavoidably, it is *pro tanto* excusing. However, the *precise nature* of the psychological deficit in question is not particularly important, since, on my view, we judge incapacity on the basis of observable evidence. For example, if a person fails the Hare psychopathy checklist, this is evidence of moral incapacity. The underlying cognitive deficit could be any number of mechanisms, which we do not need to locate. This clear emphasis on observable evidence is a part of methodological externalism, but another part, which may be still more important, is the germaneness of circumstantial excusing factors, which character theories either deny or occlude. To bring this particular difference into relief, consider three examples.

**1. Coercion**

First, consider a fairly straightforward case presented by Doris in his 2007 paper. He asks us to imagine that someone has committed perjury under coercive threat. How are we to adjudge if the person is responsible? We cannot decide on the basis of the person's choice, since the choice was strictly voluntary. That is, the agent voluntarily chose to commit perjury, albeit for the sake of some overriding concern. In order to make sense of coercion as an excuse, it seems that we must define the person's choice as 'out of character'. But how do we define 'out of character'? We have already seen that definitions appealing to statistics and relative frequency are inadequate. What possibilities remain? One option is to define 'coercion' in terms of the pressure exerted by the coercer—the intimidating mafia agent, or what have you. One of the reasons why this option makes sense is that the perjurer's psychological state is not sufficient to judge responsibility: whether the person intended to commit perjury or not, the coercive situation in and of itself mitigates responsibility by displacing some degree of responsibility onto the coercer. In this connection, Shoeman (1987) speculates that we acknowledge this distribution because we do not want to live in a society where entrapment, i.e. inducing people to commit an infraction, is tolerated. This type of explanation sheds light on the Milgram experiments as well, I believe. Suppose that, unbeknownst to us, one of the subjects had enjoyed shocking the victim. The agent would likely be blameworthy, but would he be *as* blameworthy as someone who submitted the victim to 450 volts of electricity without the inducements of the experimenter? That is, is he as blameworthy as, say, someone who built a shock machine in his basement and used it on unsuspecting victims? I do not believe that anyone would say so. This means that the agent's responsibility was partly mitigated by the experimenter's role intrinsically, i.e., regardless of the agent's intentions. This can partly be explained by reference to distributive justice. But the key point is that external factors can be inherently excusing, and so psychological accounts are incomplete.

**2. Unavoidable Evil**

For this example, consider Charles Whitman, the notorious serial killer. Whitman tragically killed 16 people. Before committing the murders, he typed a letter stating that he did not "really understand [himself]…I am supposed to be an average reasonable and intelligent young man. However, lately (I cannot recall when it started) I have been a victim of many unusual and irrational thoughts" (MacCloed online.) The letter also requested that coroners perform a post-mortem autopsy to explain his psychopathic ideation. The autopsy found that he had a brain tumor and tumor-related necrosis, which

may have influenced his behaviour (though this point is contested.) Assuming for the sake of argument that the tumor did influence his behaviour, it is fair to say that he was, to some extent, excusable. Indeed, this was the conclusion defended by Eddy Nahmias and Oisin Deery at the recent Political and Social Philosophy Conference (2014). But if character is the only condition of responsibility, then it is baffling why Whitman should be excused at all, since his tumor-induced behaviour was objectively wrong and his character was heinous. The fact that he had a tumor has no effect on the deontic status of his actions or the quality of his character, but it *does* affect his responsibility. Now, one might object that the change happened suddenly, so his behaviour can be described as 'out of character'; but if we imagine that his psychopathic behaviour had continued for years, to the extent that it had solidified into his character, then the characterological excuse fails. So character cannot be the whole story. Uncontrollable aspects of the agent's situation—in this case, the tumor—constitute extenuating circumstances.

**3. Epistemological Ignorance**

Wolf describes JoJo as an evil and sadistic dictator from a small, undeveloped country, which seems to be isolated from exogenous cultural influences. She does not clarify whether JoJo has a congenital cognitive deficit, but since we have already addressed cognitively endogenous psychopathy, let us assume for the sake of argument that he does not. He was born cognitively normal. Does it follow that he is responsible for his sadistic actions? It seems relevant in the example that JoJo, *ex hypothesi,* has never been exposed to alternative moral perspectives. If so, this must be what excuses him. Unlike the last example, this explanation does not advert to excusing aspects of JoJo's psychological architecture, since there are none. He is psychopathic, 'morally insane,' and wills his actions at the level of his deep self. Yet many people have been convinced that this sort of person is not blameworthy (viz., Wolf, Strawson, Rosen, Slote.) Wolf's explanation hinges on the fact that JoJo is excusable at the level of 'our shared pretheoretical intuitions.' Faraci and Shoemaker recently conducted an experiment to determine if naïve moral reasoners (undergraduate students) do in fact view deprived childhood circumstances as extenuating. Using their data, in chapter 7 I argue that their subjects *do* regard such circumstances as excusing to a significant degree. More particularly, there is a statistically significant discrepancy between naïve moral reasoners' judgments of evil people from ordinary circumstances, and evil people from deprived childhood circumstances. Fully resolving this question, however, would require more research, and a greater variety of examples with which to test intuitions. However, there is sufficient evidence to conclude there for many people, certain types of circumstance are significantly extenuating, irrespectively of the agent's character. Observationally, I

would say that examples clearly stating that an agent lacked moral resources from an early age generate a strong intuition that the agent is less blameworthy compared to others.

One might object that the JoJo example is too farfetched to warrant serious consideration. In real life, there simply are no JoJos, so epistemological circumstances are morally irrelevant. This is a fair objection, but it is susceptible to two kinds of response. First, one should treat the JoJo scenario as a thought experiment meant to gauge our intuitions about moral responsibility, not an actual case. By constructing an extreme example of epistemic isolation, Wolf demonstrates that such circumstances tend to be seen as excusing. In less isolating conditions, they are less excusing, but the brute *fact* of their extenuating status is established by the extreme case. The degree of their excusing force is then an empirical matter. Thought experiments enjoy a venerable history in moral philosophy as a reflective heuristic. The obvious example is trolley dilemmas, which have spawned a massive 'trolleyology.' There has never been a real-life scenario in which a person had an opportunity to stop a runaway trolley by pushing a fat man in front of its path, since no actual human is large enough to stop a trolley; but this has not prevented hundreds of people from studying this experimental setup. So, seen as a thought experiment, Wolf's example has at least as much philosophical validity as trolleyology. Moreover—and this is my second point—it is worth noting than even if there has never been a JoJo, there have been relevantly similar cases of childhood deprivation and moral insanity. I have come across such cases in autobiographical accounts of moral transformation, which are of particular interest to me for obvious reasons. Salient examples include Ishamael Beah's 'A Long Way Gone: Memoirs of a Boy Soldier' (2007), Ingo Hasselbach's 'Fuhrer Ex: Memoirs of a Former Neo-Nazi' (1996), and Sanyika Shakur's 'Monster: The Autobiography of an L.A. Gange Member' (1993). If we look at Beah, he grew up in a small village in Sierra Leone, and was forced to become a government soldier at the age of thirteen. After a process of initiation, including being forced to commit torture and take amphetamines, he committed numerous atrocities, apparently without compunction. At sixteen, he was rescued by UNICEF, expatriated to the United States, and eventually obtained a Bachelor's of political science. Now he is a political activist and UNICEF Goodwill Ambassador for Children Affected by War. In Beah's case, one can identify extenuating circumstances in the form of the commanding officer's rituals of initiation, threats, abuse, and forced drug use, which corrupted Beah's character as a youth, essentially making him, by all appearances, a psychopath. However, UNICEF managed to reform him using a variety of methods, including, by Beah's account, a policy of unconditional forgiveness. The UNICEF officials affirmed many times that Beah, as a former child soldier, was not responsible for his behaviour,

which seemed to have an ameliorating effect on him. This confirms the utilitarian advantage of such a practice, but it would not be effective in the absence of legitimate coercion. In such an absence, the behaviour would appear disingenuous, or unacceptably indulgent, and would not facilitate moral rectitude. Hence, the existence of legitimate coercion is crucial. If the same attitude were taken toward the Rwandan government, it would elicit warranted outrage, since these agents apparently made autonomous choices. This is why there must be some warrant to the claim that the excusable agent was coerced.

Hasselbach's case also exhibits extenuating circumstances (e.g., many years incarcerated in East Berlin on charges of 'rowdyism,' exposure to violence, and years in solitary confinement, which is thought to cause insanity,) but there is cause for doubt in Shakur's memoirs, which describe his life in an American gang. The issue is that Shakur ostensibly had ample opportunities to defect, which he neglected or repudiated, while he lived in a heterogeneous American community. Of course, the situation requires thorough study, but I believe that, on balance, Shakur is more blameworthy than the other two candidates in light of his circumstances. Fortunately, Beah and Hasselach were morally reformed, and Shakur renounced his gang membership, but he was subsequently arrested on several parole violations. Thus, it seems that resists moral rehabilitation, which imputes blame. The point of raising these examples is to show that one cannot conduct an appropriate analysis of moral responsibility on the basis of character alone, since all three agents were psychopaths, and yet they arguably bore different degrees of responsibility based on their personal circumstances. Character theory cannot account for this difference.

**12. Concluding Remarks**

I have argued that experiments in social psychology and cognitive science undermine internalist theories of moral responsibility and blame (such as the deep-self view) by drawing attention to (i) the epistemological problem concerning our ability to discern people's subjective motives, and (ii) the significance of the excuse of deprived childhood victims, which is not captured by internalist accounts. These considerations lend support to an externalist approach to responsibility such as SRP. This standard, which was originally implemented in jurisprudential contexts, is also appropriate for assessing moral responsibility because it balances social utility against fairness to individuals, which requires that we consider personal circumstances. Finally, in the last two sections I defended SRP against criticism, and elucidated its proper scope and relationship to normative ethics. My hope is that these arguments will prompt philosophers to reconsider their

assumptions about situationist psychology, and construe it in less deflationary terms. Properly conceived, situationism is a fairly modest and eminently useful theory, which supports an explanatory account of our moral intuitions by acknowledging the role of situational factors in moral responsibility.

## Chapter 3: Understanding SRP

### 1. Introduction

In the last chapter, I defended SRP as preferable to internalist theories such as the deep-self view and sane deep-self view. Here, I defend a particular interpretation of this standard. In legal philosophy, there are three authoritative interpretations of SRP: the indifference view, the customary view, and the avoidability view. The indifference view (henceforth, IV) was defended by M. Moran in her influential monograph on the reasonable person (2003). I believe that her approach is inadequate because it is an internalist account, which succumbs to the epistemological problems discussed in the previous chapter. In what follows, I defend the avoidability view (henceforth, AV) against Moran's criticisms. AV holds that a person is responsible for wrongdoing if, when she acted, she had the capacities required to comply with the law and a fair opportunity to exercise those capacities.

To begin my defense, I respond to a standard criticism raised by feminists, critical race theorists and critical disability scholars, which holds that SRP has been used to discriminate against vulnerable minorities, by defining 'reasonableness' by reference to the ideal of the middleclass, white male. This is a legitimate concern, which any interpretation of SRP must assuage. Moran believes that IV is the only theory capable of doing so, but I counter that this defense is beleaguered by a version of the epistemological problem. Further, I argue that misinterpretations of SRP are due not to any intrinsic and unique defect of AV, but to an epistemic defect in judges, which can afflict the application of any principle. Thus, IV has no advantage in this regard. This discretionary problem can be ameliorated using procedural constraints on judicial appointments, which facilitate epistemically virtuous verdicts. Moreover, I argue that IV inadvertently discriminates against people with psychological disorders, since it dismisses normative-cognitive deficits. Overall, I show that AV manages to avoid the epistemological problem and to accommodate widely shared intuitions about epistemic isolation and neglect better than any of the alternatives.

After elaborating these points, I outline how SRP can be "particularized" (Moran, p. 281), i.e., fleshed out in precise terms, to accommodate the interests of women, racialized minorities, and disabled persons, who were excluded from Holmes' original standard 'of the reasonable man.' The courts have established precedents for new, epistemologically-responsible

interpretations of 'reasonable personhood,' stemming from a multitude of interdisciplinary work, including social and political philosophy, feminist philosophy, cultural anthropology, and expert testimony in court proceedings. I briefly outline these precedents and suggest strategies for fostering greater sensitivity to morally relevant information in judicial deliberation, in the interest of protecting the interests and constitutional rights of political minorities. This includes introducing procedural constraints such as appointing diverse individuals to positions of legal power.

**2. M. Moran and the Indifference View**

SRP plays a central role in determining culpable negligence in criminal and tort law. However, in the history of judicial proceedings, it has unfortunately been applied in such as way as to discriminate against women, racial minorities, and disabled persons. Moran proposes to solve this problem by defining the 'reasonable person' so as to include certain biographical or empirical features of the defendant (such as age, intelligence, level of education, and physical abilities,) but to exclude specifically "prudential or normative shortcomings" (p. 243). On her view, a defendant is liable for negligence if she "fails to notice a risk because she is indifferent to the interests of others" (2003, p. 258), but not if she lacks other kinds of capacities. This view is meant to "capture the normative failing betrayed in culpable inadvertence" (Moran, p. 257).

I shall argue that this account is inadequate on two scores. First, it requires that we ascertain an agent's relevant mental states with reasonable certainty, which is one of the difficulties that Holmes' account was specifically designed to avoid; and second, it requires that we draw a reliable distinction between *cognitive shortcomings* (such as intellectual disabilities) and *normative or prudential shortcomings* (such as sociopathic personality disorder), which may be impossible in practice. In what follows, I contend that AV, which relies on a testable, comparative notion of avoidability, bypasses these epistemological problems, and also protects the interests and constitutional rights of political minorities as any general theory can. I defend this position by appealing to persuasive epistemic accounts of moral agency (viz., Fricker 2007, Rosen 2003, Slote 2002), and the Harvard Law Review Association's (1986) argument for a defense of culture. These arguments support AV using a variety of argumentative means. Finally, I outline how we can refine SRP to protect the interests of vulnerable minorities, though one cannot expect to eliminate personal discretion from legal and moral reasoning. Thus, some degree of practical wisdom is needed.

**3. Clarifying SRP**

Holmes defends SRP on grounds that we saw in the last chapter. Unfortunately, in spite of its theoretical merits, SRP has been applied in a discriminatory fashion. Moran believes that its dependence on prevailing notions of "normalcy" and "naturalness" tends to rationalize "the

unequal treatment of both the developmentally disabled and women" (p. 147). For example, SRP has been used to seek the involuntary non-therapeutic sterilization of persons with intellectual disabilities, to exclude women from professional and public life, and to subject women to harsher sentences than men (Moran, pp. 147-48)—clearly atrocious results that we should condemn. But are these implications attributable to the AV *per se* (assuming that this particularly interpretation informed the condemnable judicial decisions)? Moran's proposal is to disregard the empirical or biographical aspects of the litigant, and narrow in on "the normative core of the standard"—the agent's intentions (p. 260). This view echoes R. A. Duff's classic 'indifference account,' which triggers liability if a defendant's actions displayed "a seriously culpable practical indifference to the interests which her action in fact threatened" (Moran, p. 260). This approach, says Moran, allows us to differentiate between "prudential or normative shortcomings," which are blameworthy, and "those shortcomings (like cognitive and physical ones) that do not reflect upon the character of the agent" (p. 243). For instance, it allows us to excuse a driver who kills a pedestrian due to an unforeseeable medical emergency, and sentence someone who drives recklessly in traffic due to indifference.

Although IV purports to answer the concerns raised by equal rights advocates, it is not immune from criticism. In the next section, I adduce two objections which I take to be particularly convincing. Then, in section 4, I explain why Moran's objections against AV are not decisive.

**4. The Indifference View: Two Criticisms**

In order to see why Moran's objection is not fatal, it is important to understand why SRP, as an *objective theory*, was originally developed by Holmes, and why it came to be favoured as one of the predominant methods for determining responsibility, especially in cases of negligence, recklessness, coercion, and necessity.

An *objective* theory is one that ignores a defendant's *de facto* subjective states and instead considers what a reasonable person would have done in relevantly similar circumstances, given relevantly similar capacities. This hypothetical construct may take into account certain empirical or biographical aspects of the defendant, such as age, physical ability, and whether the agent was facing an emergency. In addition, judges have begun appealing to the agent's social and epistemic standing, as in R. v. Lavalee (1990).

In contrast to SRP, the *subjective* approach is based on an investigation of the defendant's subjective mental states, as expressed in her individual choices or settled character traits. Accordingly, there are two standard subjective theories: *choice theory* and *character theory*. Choice theory, recall, holds an agent responsible if she broke the law by an action that was a result of her free choices, and character theory deems an agent responsible if she broke the law by an action that was the expression of her settled character, or robust mental

dispositions. As we saw in chapter 2, subjective accounts are vulnerable to the *epistemological problem*—the problem of poor introspective access. IV, on scrutiny, is a subjective account.

Notably, Moran takes pains to clarify that her view is not supposed to be seen as a form of character theory, due to well-know complications with this approach. She says that although IV ostensibly appeals to character, insofar as it demands that the litigant defend herself from the charge of criminal character, it nonetheless does not define 'indifference' in characterological terms:

> Although this task of justification does implicate "character," it does so in quite a different way than in character-based theories of culpability. Unlike at least some versions of character theory that inquire into the general or stable character and attitudes of the agent, an indifference account focuses on the character of the agent's choices and deliberations in posing the relevant risk. The question therefore is whether the choices and actions that gave rise to her mistake betray her indifference toward others. (p. 263)

In highlighting the agent's *choices* and *actions* as opposed to character, Moran aims to avoid traditional objections to character theory, which she recognizes as especially forceful. However, in doing so, she inadvertently opens herself to these very objections, by espousing a form of choice theory. As Doris points out, choice theory "slouches" into character theory, because choices made under duress must be defined as out-of-character in order to deflect culpability, and so free choices must be defined in terms of the agent's "rational judgment," "evaluative commitments" or "plan of life"—i.e., the agent's settled character (Doris, p. 521). But character theory is susceptible to two additional objections: (i) that experiments in social psychology impugn the existence of settled character, and (ii) that there is no reliable means of differentiating 'in character' behaviour from 'out of character' behaviour, which is part of what Doris calls *the epistemological problem*. This problem, we have seen, calls for the adoption of an objective account (such as AV.)

To briefly reiterate the particulars of this problem, we do not have reliable access to another person's psychological states, in light of psychological inscrutability, internal fragmentation, confabulation, rationalization, and the like. Moreover, neuroscientific research indicates that human beings are highly susceptible to cognitive biases, distortions, and illusions, which hamper our ability to discern our own motives. In sum: (1) Gazzaniga et al.'s split-brain tests (1977) suggest that we may be confabulating more often than we think; (2) Libet's studies on (1992) unconscious neural processes, which seem to preempt volitional responses, suggest that we may very often confabulate explanations for previous 'decisions' that were actually unconscious responses; and (3) Zimbardo's social science research supports the inference that "people are less rational than they are adept at *rationalizing* – explaining away discrepancies

between their private morality and actions contrary to it" (2007, p. 220), and rationalization tends to distort self-perception.

A second type of problem concerns Moran's claim that an agent can be held responsible only for "a prudential or normative shortcoming" (manifested as "culpable practical indifference to the interests of others"), but not a "cognitive shortcoming," which putatively "does not reflect upon the character of the agent" (Moran's terms, p. 243). If the epistemological problem is correct, then it is not clear how we are to make the postulated distinction. Moreover, cognitive shortcomings and prudential/normative shortcomings are often functionally imbricated and immiscible—for instance, when an agent expresses indifference (a normative shortcoming) due to congenital psychopathy (a cognitive deficit.) Is the agent's indifference ultimately normative or cognitive? Seemingly both. A congenital brain deficit is surely a form of *cognitive deficit*, since it does not arise from any voluntary choice on the part of the agent—it is simply present at birth; yet it can easily give rise to normative deficits, such as psychopathy, or some other moral-psychological disorders (categorized under 'personality disorders' in DSM 5.) How are we to distinguish between these two types of deficits, one 'cognitive' and the other 'normative'? On this point, Poythress and Skeem (2006) show that psychopathy can result from either "a congenital affective deficit," or an "acquired, environmentally based affective disturbance" (p. 174). This latter type of deficit is potentially 'normative,' in Moran's sense of the word, but it is scientifically and symptomatically indistinguishable from the former 'cognitive' defect, which emerges spontaneously in the brain. The researchers' work draws a *conceptual* line between *cognitive* and *normative*, but they emphasize that there is *no practical way* of differentiating the two deficits. They note that although there are promising strategies for disambiguating etiological pathways in the future, "research lags far behind clinical and speculative musings" (p. 175). That is, there is currently no way of distinguishing cognitive from acquired (normative) deficits. A further complication is that psychopathy is believed to be a dimensional, as opposed to categorical, condition, meaning that there is not clear line between mental illness and voluntary choice: the two blur into one another. This implies in turn that a cognitive-versus-normative distinction, which is supposed to inform moral judgment on IA, is simply impracticable. Hence, it cannot inform practical judgments of legal (or moral) responsibility. By contrast, AV, as I have defined it, eschews strict psychological data and relies on overt evidence.

**5. Alternatives: Defending an Avoidability Interpretation**

Moran considers, but ultimately dismisses, two alternative accounts of SRP, which we are now in a position to evaluate in some detail. The first is the *customary account*, which defines responsibility in terms of social customs—specifically, in terms of what the community would expect of an average person. The second is AV, popularized by the legal scholar H. L. A. Hart, which requires as a precondition for responsibility that the agent had "the normal capacities,

physical and mental, for doing what the law requires and abstaining from what it forbids, and a fair opportunity to exercise those capacities" (Hart 1970, p. 152). Moran rejects both accounts on essentially the same grounds: namely, because although they may "coincidentally embody respect for equal moral standing, [they] will not *necessarily* do so" (2003, p. 247). That is, they may simply reinforce customary discrimination, which contravenes one of the main functions of the justice system.

It is worth considering the relationship between AV and the customary view. The two overlap, inasmuch as customary behaviour naturally and inevitably informs our notion of what one may expect of a reasonable person. Doris points out, for example, that AV is informed by population base rates, given that, "surely it is partly because most people yield under torture that it seems unfair to hold victims liable for failing to resist it (p. 527). On the other hand, there must be allowances for saying that certain practices or patterns of behaviour are blameworthy *even though* they are widespread. Otherwise the law cannot fulfill its purpose of enhancing security and promoting responsible citizenship. Doris does not, in my opinion, give particularly clear guidelines on how to draw this distinction, but there are helpful resources in feminist epistemology. Philosophers such as Fricker have argued that morally responsibility requires not only (i) the capacities required to comply with particular norms and a fair opportunity to exercise them, but also (ii) access to the shared tools of social interpretation, which are required to understand and appreciate such norms, and (iii) a fair opportunity to *internalize* such norms by incorporating them into one's daily life. Accordingly, Fricker observes that, "one cannot be blamed for making a routine moral judgment. But one can none the less be held responsible for making a merely routine judgment in a context in which a more exceptional alternative is, as a matter of historical possibility, just around the corner" (p. 105). This is why she holds Greenleaf partly excusable for doubting Marge. In the same vein, G. Rosen (2003) contends that the Hittites can be excused for practicing slavery because, "given the intellectual and cultural resources available to a second millennium Hittite lord, it would have taken a moral genius to see through to the wrongness of chattel slavery" (p. 66). And M. Slote (1982) claims that ancient slaveholders can be excused for their actions because they were "unable to see what virtue required in regard to slavery…, not due to personal limitations (alone) but [also by virtue of] social and historical forces, by cultural limitations" (p. 72). Notably, these philosophers agree that this type of excuse would not apply to ordinary citizens of modern democratic societies, who were socialized into the prevailing customs through standard interpersonal processes. Hence, this view does not threaten the authority of the relevant norms (whether legal or moral.) Thus, it does not threaten the integrity of the principle of *ignorantia juis non excusat* (ignorance of the law does not excuse), since ignorance would only excuse under rare epistemic circumstances.

Although these are accounts of *moral* responsibility, the same kinds of consideration translate into the domain of legal responsibility. Indeed, many legal philosophers believe that

criminal law aims to classify kinds of *moral wrong* and to punish those who commit them, which collapses the distinction between legal and moral reasoning (e.g., Moore 1997, Tadros 2011.) But even if one does not subscribe to this view, legal scholars such as Dworkin (1977) maintain that 'thin' moral concepts at least—such as fairness and social welfare—inform judicial decision-making. These views regard moral and legal responsibility as permeable. Another important consideration is that moral and legal *permissibility* (which the former accounts address) are different from moral and legal *responsibility*, and while the former may diverge quite drastically, the second may coalesce. This would be true if an agent's responsibility for violating a given norm depended on her circumstances and capacities, whether the given norm is moral or legal. This is my view, and also resembles Peter Cane's account in 'The Nature and Function of Responsibilty (2002). This position will be elaborated in chapter 4, but for now I will pursue the defense of AV.

One area in which avoidability plays a significant role is the legal *defense of culture*. As The Harvard Review Association (HLRA, 1986) points out, in American criminal law,[6] "it may be possible in some cases to introduce cultural factors into court under the rubric of the insanity defense"; and, "under existing legal doctrine, cultural factors may be relevant insofar as they negate the intent required to satisfy the definition of a crime" (p. 1294). Furthermore, prosecutors and judges may take cultural factors into account when charging a defendant with a crime, plea bargaining with a defense attorney, and determining appropriate sentencing. These are informal discretionary procedures, but there are philosophical grounds for introducing a "formal, substantive defense of culture" (HLTA, p. 1296). In particular, in a fair system of justice, a person should receive a mitigated sentence if she: (a) "committed a criminal act solely because she was ignorant of the applicable law" (p. 1299), or (b) "committed a criminal act solely because the values of her native culture compelled her to do so" (p. 1300). The first condition recognizes the legal relevance of unavoidable ignorance, and the second condition appreciates that human beings require time to grasp and internalize new customs, as a matter of psychological fact. To ignore these facts would be to place ingénues under the same responsibility burdens as nationals, which we be unfair in light of their circumstances.

These examples provide guidelines for how to apply AV without jeopardizing the deterrent effect of public sanctions. HLRA contends that the defense of culture does not threaten the public good, as it can only properly be invoked in "extraordinary circumstances that are unlikely to recur," or else when the defendant acted "solely from ignorance of the law" in which case "punishment is unnecessary to deter similar conduct in the future" (p. 1303). Similarly, Fricker's view has fairly limiting exculpating potential, given that it is meant to excuse ignorant behaviour only when, and to the extent that, a more enlightened perspectives was unreachable

[6] Essentially the same observations apply to Canadian law.

"as a matter of historical [or cultural] possibility" (p. 105). These excuses clearly are not meant to apply to most people in liberal democracies, so AV is not a threat to the common good.

**6. Equality Concerns**

Moran's main objection to AV is that it incites discrimination against political minorities. Specifically, she says that the concept of avoidablity is too vague and contestable to be applied responsibly and fairly, especially in cases of persons with learning disabilities who have historically been treated as if non-disabled. My response to this charge, in essence, is that the misapplication of SRP is not due to any inherent feature of AV, but is attributable to the biases and partialities of judges, which would affect the application of any principle. As a *theoretical* approach, AV view is no more vulnerable to this criticism than IV.

If we reflect on the history of the law, it is fair to assume that SRP was generally misapplied because prosecutors and judges misunderstood what it would be reasonable to expect of particular defendants under different circumstances. However, this problem is gradually receding due to increasingly progressive legal precedents, broader representation in the judiciary, enlightened legal education, employment equity, expert testimony, and the like. We saw this in R. v. Lavalee. The factors that made a difference to the case, I believe, were the introduction of expert testimony on the effects of misogynistic violence, and the enlightened perspectives of the judges. Using these means in future trials can increase responsible judicial discretion, regardless of the legal principle under consideration. However, Moran's view can be seen as susceptible to a distinctive and irremediable type of misapplication, since it postulates a distinction between innocent cognitive deficits and blameworthy prudential or normative shortcomings. This elision prevents us from excusing cognitive-normative mental disorders such as insanity, although this excuse has been effective in Canada since the 1990s and in Australia since the 1930s, and serves to protect defendants from unreasonable and ineffectual retributive sanctions. Moran's view entails rescinding this provision, since it defines indifference as culpable regardless of the underlying cause. Surprising, the revisionary implications of IV for legal practice are nowhere explicitly addressed by Moran. Yet if she supports this revision, she must show that persons with mental disorders should be subject, without exception, to the same sanctions as psychologically normal persons, even though their motivational structure may be appreciably different. Since she nowhere speaks to this issue, her account is at best incomplete.

I believe that a better solution to judicial error and epistemic ignorance lies in introducing and reinforcing procedural safeguards, such as those defended by David O. Sears (1994) and Keith C. Culver (2008). Very briefly in the time remaining, these authors suggest that we should attend "to history and social science for evidence that law, legal concepts, and legal institutions contain biases against certain groups or interests" (Culver, p. 212); note "the substantive effect of

law more than its formal appearance" (Culver, p. 213); appoint more political minorities to the bench to mitigate implicit prejudice; familiarize law students with alternative perspectives such as feminist jurisprudence, critical race theory, and critical disability theory; and incorporate more expert testimony from a variety of sources into the courtroom. Unlike Moran's proposal, these safeguards do not place absolute constraints on judicial discretion, which is arguably indispensable for considering the unique features of each case and appropriately balancing all factors.

### 7. Particularizing the Standard

This still leaves the question of how to particularize (or conceive of) SRP, in such a way that we can excuse persons with innocent intellectual and cognitive disabilities, but blame persons with patently blameworthy failings. Moran highlights this need when she observes that, "virtually the only place where the rigid reasonable person rule is adhered to in the face of significant tension with the underlying principle of avoidability is in the case of people with developmental disabilities" (2003, p. 176). She proposes IV as a solution. However, it is significant that her *own description* of misapplications of SRP describes them as being in 'significant tension' with the principle of avoidability, rather than in accordance with it. This implies that a *proper understanding* of avoidability rules out such applications. It follows that AV is not inherently flawed, but contingently misunderstood, which lends credence to my thesis and proposed solutions.

Nonetheless, Moran suggests that only IV is able to ensure that, "at a minimum, the reasonable person [is] understood as committed to fundamental constitutional values" (p. 284). But this is a hasty conclusion. AV surely also supports this commitment, at least in modern democratic countries that have publically-accessible constitutions. (One can find a copy of the Charter of Rights and Freedoms on the Government of Canada website.[7]) In this climate, we expect a reasonable person to be at least somewhat familiar with basic constitutional principles, and to act accordingly, unless there are rare extenuating circumstances such. So AV captures the constitutional requirement, in effect. However, unlike IV, it also acknowledges the relevance of epistemic factors as potential excusing factors. These factors explain why, in standard legal practice, there are grounds for partially excusing recent landed immigrants and insane defendants. When contemplating these types of excuses, Moran endorses D. Archard's (1999) assertion that "everyone has a fair opportunity to develop a non-defective character" (Moran, p. 268). But it seems very implausible to say that *absolutely everyone* has internalized cultural standards to exactly the same degree, whether born in the country in question or on the other side of the world. It is simply unreasonable to make a sweeping, *a priori* claim about epistemic

[7] http://laws-lois.justice.gc.ca/eng/const/page-15.html

access, without considering each person's situation. Surely it makes more sense to say that people have differing capacities to appreciate the law based on their differing epistemic circumstances, the latter of which is a matter of empirical inquiry. Indeed, Archand's claim practically implies that everyone is omniscient regarding legal standards, since he holds everyone to the same bar.

Now to the main question: how should we understand AV? Although this cannot be determined *a priori,* or through conceptual analysis alone, we can gain insight by considering landmark cases that challenged historical understandings. There are four special interest groups which may require specialized interpretations of avoidability: women, cultural minorities, people with physical disabilities and/or medical problems, and people with mental health disabilities. I shall address the first three here, and reserve the fourth for chapter 5, which addresses mental health disabilities specifically.

1. We saw in R v Lavallee that the court applied SRP to excuse the defendant. Another area where this standard has been utilized extensively is in sexual assault cases. Perry et al. (2004) note that the "reasonable woman standard" is invoked to define a "hostile work environment" from a woman's perspective, and that "one justification for using a reasonable woman standard is the large body of social science research that suggests that men and women perceive harassment differently (see Blumenthal, 1998 and Rotundo, Nguyen, & Sackett, 2001 for meta-analyses of this literature)" (p. 11). These results "suggest that hostile environment sexual harassment cases that operated under a reasonable woman standard were somewhat more likely to be decided for the plaintiff than cases that did not operate under a reasonable woman standard" (2004, p. 22). Thus, there is some observed benefit to using this standard in practice. In this respect, gender might be relevant to judging avoidability.

2. Cultural accommodations are less common, although there are germane examples. In one well-known case (1985), a Japanese woman who had resided in America for more than a decade drowned her two young children and then tried to commit suicide in a Japanese cultural practice called *Kimura* after learning of her husband's adultery. The prosecutor did not excuse the defendant, but he did accept her plea bargain to the lesser charge of voluntary manslaughter. This was partly based on the consideration that "one could expect a Japanese [woman] to react in a much more volatile, violent way to those circumstances than someone from our own [American] society" (HLRA, p. 108). This verdict aligns with the avoidability view, insofar as, on the one hand, the defendant was not well-integrated into mainstream society, but on the other hand, since she had been

living in American society for a decade and had opportunities to observe local customs, she does not deserve full exemption. This is consistent with the judge's verdict. This example provides rough guidelines for how avoidablity can be applied in cases of cultural isolation.

3. Finally, the avoidability view captures how medical conditions or disabilities can be excusing in court. This is illustrated in two driving cases which produced to two different verdicts. In the case of Roberts v Ramsbottom (1980), "the defendant who had had a stroke was held liable for an accident he caused despite the fact that the court accepted that the stoke rendered him 'unable to appreciate that he should have stopped'" (Moran, p. 22), whereas in Mansfield v Wheatabix (1998), "the Court of appeal refused to find liability on the part of a man who partially lost consciousness as a result of a hypoglycemic state cause by a serious malignancy he did not know he had" (Moran, p. 22). The judge noted that Ramsbottom, unlike Wheatabix, knew that he was ill, and therefore should have pulled to the side of the road when he began experiencing symptoms. He had the opportunity to do so, but chose not to. Thus, he was deemed liable and sentenced. This indicates that people should be blamed for harm due to disabilities only if they knew or suspected that the disability would be a threat to others, and did not act on this knowledge.

These examples indicate how an AV reading of SRP can be applied to specific cases. In tailoring SRP, we must try, to the best of our ability, to understand and appreciate the epistemological spaces of women, cultural minorities, and persons with health conditions and disabilities. Attending to relevant legal precedents such as those cited above can help us to achieve the desired epistemological attunement, but we must also consistently try to hone our capacity for what Fricker calls *epistemic virtue*, in order to fully appreciate and accommodate alternative perspectives. Fricker provides an excellent account of this practice, which I will outline here. She begins by delineating two kinds of epistemic injustice: testimonial injustice, whereby one underestimates a speaker's credibility, and hermeneutical injustice, whereby one dismiss a speaker's testimony due to a paucity of interpretive (hermeneutical) resources. In the second case, a person may not be able fully to appreciate her own position of oppression, but she is still subject to substantial harms (such as a lack of rights, opportunities, respect, etc.) We are all susceptible to these distortions, but we are morally required to attend to them and try to reverse them, in order to enhance our epistemic virtue. Fricker is somewhat obscure on the means of achieving this, but they involve, at least, reflective deliberation, and taking measures to ensure a more equal distribution of epistemic resources. The first is a (mainly) personal endeavour, and the second is a (mainly) political one, which requires structural change. It is interesting that these

solutions mirror a proposal made by Doris and Merritt in an article on character (2010), which contends that the situationist psychology findings imply that we ought to foster stable environments by (a) enhancing practical reasoning, and (b) creating "regularly recurring social contexts," using legal and social regulatory mechanisms (political measures, legal enforcement, and so on) (p. 390). Elsewhere, I have strongly endorsed (b), and added that, due to the nature of the epistemological problem, we should foster interpersonal relationships that promote (narrow) character-development via *habituation*, which implicates system-1 (affective) processes. Because this system utilizes sub-cognitive, emotional mechanisms (realized primarily in the reward system and basal ganglia,) it should be less susceptible to cognitive distortions than 'system 2' computational processes. This analysis supplements Fricker's account of virtue epistemology by identifying additional means. The upshot is that if we wish to employ SRP responsibly, we should cultivate epistemic virtue through vigilance, social policy, and sensitization. Hence, although the above examples may be instructive, individuals must develop epistemic virtue to apply AV responsibly case-by-case. Some scenarios will be not obviously resemble previous cases.

**8. Conclusion**

I have argued that Moran's defense of IV is called into question by the epistemological problem articulated by Doris, and the difficulty of disambiguiating normative shortcomings from cognitive ones as a precondition for ascribing responsibility. Furthermore, AV provides a stronger theoretical foundation for standard legal excuses, based on Holmes' positive account, together with intuitive examples in moral philosophy, feminist epistemology, and criminal jurisprudence. While Moran's view addresses a legitimate problem, it is a form of character theory, which even Moran admits to be problematic; it requires epistemic virtue for correct implementation, just like any other general theory; and it excludes the mental disorders defense without argument. To remediate prejudiced judicial decision, I have recommended procedural constraints.

These arguments support the use of SRP for judging legal responsibility, but I believe that they extend to deliberation about moral responsibility as well. This is because, although there may be significant differences between moral *permissibility* and legal *permissibility,* responsibility in both domains is a function of whether a particular transgression was *avoidable* by a particular person in particular circumstances. Thus, whether a person has committed a legal infraction (e.g., murder) or a moral infraction (e.g., willfully malicious adultery), the person's responsibility can be determined by SRP. I defend this position in the next section.

**Chapter 4: Legal Versus Moral Responsibility**

In this chapter, I explain the relationship between legal responsibility and moral responsibility, which I have described as permeable. Yet because this is a contentious position, it requires

considerable argumentation.

First, it might be helpful to consider the relationship between legal theory and moral theory, which are broader topics. (Moral responsibility is only a small subset of moral theory, which comprises normative ethics, metaethics, moral psychology, moral epistemology, and so on and so forth.) What marks the boundary between legal theory and moral theory? There is no consensus on this point, but there are three well-known views on the subject. The oldest account, natural law theory, holds that there is no strict division between the law and morality. Laws are natural artifacts or natural kinds, determined by certain facts about human beings and the world. Some theorists (such as Aquinas) believe that natural laws emanate from God's will, and others believe that they derive from human nature (viz., J. Finnis 1980), human capacity, human rationality, human biology, or some other feature of human beings. These theorists make a distinction between real laws, which cohere with morality, and ersatz laws, which diverge from morality. (We might envision laws as a subset of morality, being justified on the same basis as moral principles, i.e., human nature.) Gustav Radbruch, who converted to natural law theory after the Holocaust, argued that immoral laws are no laws at all, and so Nazi 'laws' were never authoritative. This licenses the post-facto conviction of Nazi officers and collaborators, on grounds that they violated natural laws against human rights violations, which were not given appropriate recognition under the Nazi 'legal' system (so-called.) The irony in this account, which Hart emphasizes in his well-known critique (1958), is that it makes it impossible to say that there are any unjust laws. At most, we can say that there are unjust law-facsimiles, which are falsely entrenched in quasi-legal systems. But the natural inclination is to say that some laws are immoral. Being able to say this also provides a motive to reform laws through institutional means.

The second standard theory, which was once dominant but is "no longer clearly the leading theory of law" (Culver 2008, p. 75), is legal positivism, which posits a necessary distinction between law and morality. The foremost protagonist of this view is Hart, but his view is "increasingly unpopular" (Culver, p. 75). This account is the most important to consider from my perspective, because I have posited a connection between moral responsibility and legal responsibility—though I have maintained that it differs from the relationship between moral theory and legal theory, which I see as concerned primarily with permissibility, obligation, or virtue (depending on one's theoretical commitments.) Hart believes that we must distinguish between law and morality for purposes of clarity: as Culver puts it, "by recognizing that laws can be used for many purposes, both good and bad, we gain a clearer picture of the possibilities and limits of law as a tool for guiding behaviour" (p. 76). This distinction is also useful because it motivates us to reform laws that are ostensibly immoral, as opposed to merely dismissing them and acting as if they did not exist. There would be anarchy if people only obeyed laws that they deemed moral on their subjective understanding. In spite of endorsing this formal dichotomy, Hart concedes that laws and morality intersect at certain junctures. First, he admits that, "as a matter of historical

fact, the development of legal systems had been powerfully influenced by moral opinion, and, conversely, that moral standards had been profoundly influenced by law, so that the content of many legal rules mirrored moral rules or principles" (p. 598). So he acknowledges a contingent, historical connection between laws and morality. Secondly, he admits that morality can be expressed through national constitutions, which place constraints on the scope of the law. Thirdly, he believes that morality may constrain the legitimacy of a *legal system*, though not individual laws: "Perhaps the differences with respect to laws taken separately and a legal system as a whole are also true of the connection between moral (or some other) conceptions of what law ought to be and law in this wider sense" (p. 621). Certain fundamental laws, such as those against murder, violence, and theft, he says, are grounded in human nature, but others do not have this "necessary nonarbitrary status" (p. 623). Finally, Hart says that there are "penumbral" cases which require judicial discretion, and in such situations judges should consider the social consequences of their decision. For example, if the law states that a person may not take a stolen 'vehicle' across state lines, and someone flies an airplane across the border, the judge must determine if the airplane satisfies the definition of a 'vehicle.' In doing so, she should consider whether her interpretation advances certain acknowledged social goods. Insofar as the social good is a moral reason, judicial discretion may be seen as morally permeated.

The third standard theory is Dworkin's integrity view (1977), which disputes Hart on the point of judicial discretion. Dworkin contends that Hart's account misses an important part of the discretionary process, which is the role of principles. Judges cannot decide penumbral cases on any basis willy-nilly, but must appeal to principles of justice and fairness. These principles exist independently of considerations of social utility, on their own merits: "I call a 'principle' a standard that is to be observed, not because it will advance or secure an economic, political, or social situation deemed desirable, but because it is a requirement of justice or fairness or some other dimension of morality" (Dworkin, p. 154). According to Dworkin, this revision undermines the moral-legal distinction, and refutes positivism. This objection has significantly influenced current legal debate, spawning more moderate forms of positivism and alternative accounts.

My personal inclination is closer to Dworkin's than to Hart's, but these examples show that *all* legal theories are, in some regards, connected to morality. Natural law theory sees the two as identical, or at least isomorphic; integrity theory sees them as overlapping in the ambit of judicial discretion and legal reform; and even legal positivism admits several significant points of intersection. So it would be false to assume that law and morality are completely divorced, except on a very radical and obscure conception of law. That said, it would also be false to think that the two frameworks are coextensive. Even if natural law theorists believe that law and morality derive their authority from the same source, i.e., human nature, they admit that not all moral infractions deserve legal status. For example, adultery is not worthy of criminal sanctions. The reasons for this are manifold, but one salient rationale is that there should be some degree of separation

between private and public life, which permits people the freedom to pursue their own projects and interests. In spite of this distinction, however, there is general consensus that the legal system, however circumscribed, must comport to some conception of objective morality or some conception of the social good.

For the purposes of my project, I do not need to commit myself to any specific conception of law (natural law theory, positivism, or integrity view) or any particular ethical theory (Kantianism, consequentialism, virtue ethics, or contractualism,) because I am not proposing a normative ethical theory. Rather, I am proposing a theory of moral responsibility, which has distinctive features, and can be specified relative to *any* given normative framework. Therefore, if, say, Kantianism turns out to be true, and utilitarianism false, my theory of responsibility still holds. The viability of any given ethical theory has no significance for my account of responsibility, so long as some moral facts are true—that is, so long as nihilism is false. This is the case for almost all theories of moral responsibility. Frankfurt's account, for example, will not turn out to be false if utilitarianism is false, or if Kantianism is false, since it is distinct from any overarching ethical framework. All that Frankfurt's theory requires is the possibility of the formation of second-order volitions, and fairly good introspective access to one's psychological economy (which I have disputed.) Likewise, all that my theory requires is information about the agent's interpersonal capacities, personal history, and circumstances. If Kantianism is true, and a truly morally worthy action conforms to the categorical imperative, then my account of responsibility will entail that *if* someone fails to conform to this principle, then the person is presumptively morally responsible, unless and to the extent that her action was unavoidable in the sense already specified. *Mutatis mutandis* for utilitarianism, virtue ethics, and contractualism.

To bring this point into relief, consider the case of adultery. Adultery is not a legal offense in the developed world, presumably because it is too mild an infraction to warrant criminal sanctions. Is it a moral offense? According to act utilitarianism, an act that maximizes utility is not only morally permissible, but obligatory. So if an act of adultery maximizes utility, it is obligatory. According to rule utilitarianism, an act that conforms to a utility-maximizing rule is morally obligatory. So adultery is obligatory if it conforms to such a rule. This may be the case for certain instances of adultery—for example, if one's partner is in a coma, or, according to D. Marquis (2005), if one's partner denies one sexual access. (I do not agree with this argument, but it illustrates the hermeneutical possibilities.) According to Kant, adultery is wrong if it cannot be rationally universalized. Kant believed that it could not, but he also proscribed marital sex without the possibility of procreation—for instance, for same-sex couples and infertile couples (MM, Ak 6:277–79, 6:424–427), as well as premarital sex and masturbation. Most contemporary Kantians would probably want to construe the categorical imperative in a more progressive light, but there are manifold extant interpretations of Kantianism, which cannot be given here. The point is that it is a difficult question. According to Aristotelian virtue ethics, adultery is wrong if it expresses a

vicious motive, so the agent's motives are crucial. This, of course, does not give us an answer to whether any *particular* instance of adultery is wrong, but virtue ethicists are quite comfortable with the fact that virtue ethics does not offer cut-and-dried solutions. They acknowledge that judging the virtue of an action requires practical wisdom, as well as non-cognitive (affect, intuitive, perceptual) sensitivities that must be developed over years of experience (viz., McDowell 1979, Hursthouse 1999). So we do not have a clear answer as to whether adultery *tout court* is morally wrong, nor do I purport to be able to solve this quandary. Fortunately, my account of moral responsibility does not require that we select any particular ethical theory, or resolve any particular ethical dilemma. It comes into effect only after we have *already deduced* that an act is wrong. Then and only then can we begin to inquire into whether the agent was also *responsible* for the putative moral transgression, by applying SRP. In this way, deciding questions of moral permissibility (or obligatoriness, or viciousness, depending on one's theoretical orientation) is logically prior to deciding questions of moral responsibility. To simplify matters, suppose that someone has committed an ill-motivated, unprovoked act of adultery, which confers no foreseeable social benefit. All three previously-discussed ethical theories converge on the conclusion that *this* act of adultery was morally wrong (in some of the word.) Only now that we have resolved the broader normative question can we inquire into whether the agent was responsible for the given moral transgression. Presumptively, the answer may seem to be yes, but we must scrutinize the agent's circumstances to see if an excuse is forthcoming. Provided the agent is psychologically normal, had a normal course of development, and lives in a relatively free and open society, she is responsible. Other accounts of moral responsibility, such as Frankfurt's, likewise describe responsibility in *ex ante* terms, relative to actions that have already been deemed wrong or impermissible. This seems to be a methodological constraint on most theories of responsibility (viz., Frankfurt, Taylor, Watson, Wolf), although Smart (1961) offers an explicitly utilitarian account, and Scanlon (1998) gives a contractualist gloss, although on this topic, he nowhere appeals to principles that no one could reasonably reject, though he does appeal to an agent's ability to "avoid a burden" (249). So while Scanlon is a contractualist regarding moral obligation (what we owe to each other), his account of responsibility hinges on avoidabilty.

This points to the central difference between permissibility and responsibility in my view, i.e., that the latter rests more heavily on avoidabilty (or ought-implies-can.) This is clear in 'naturalistic' Kantian accounts, which interpret Kant's assertion, "What one ought to do, that one can do," in pragmatic terms. 'Idealistic' versions of Kantianism, in contrast, deny this principle (Brown 1950). They maintain that, "an increase in the number of failures to discharge a duty strengthens the ground for condemning the agent as immoral" (Brown, p. 276). But I do not think that this counts against my reading, for this refers to the agent-based, attributive sense of 'responsibility,' not the 'accountability' sense of responsibility being discussed here. We can,

without contradiction, regard multiple moral infractions by an agent as imputing invidious character, as well as indicating a lack of responsibility. (This explains extenuating judgments of psychopathy, for instance.) Moreover, Brown's account of Kantianism distinguishes between *obligation* and *punishment*—which mirrors my distinction between permissibility and responsibility—and deems punishment, uniquely, to depend upon circumstantial factors such as "the agent's glands, his parents, his neighbourhood, and his subconscious" (p. 278). Also notably, Brown says that these factors are particularly relevant in light of social science research, which increasingly discloses exculpating factors. Thus, in spite of initial appearances, Brown's view complements my own quite well, since his theory of punishment corroborates my account of responsibility *qua* deployment of the negative reactive attitudes. This account of responsibility enshrines avoidability as a central principle, requisite to determining appropriate sanctions.

Other Kantians, too, typically accept some version of ought-implies-can, on the following general schema: "If anyone cannot perform an act, he is not required (i.e., responsible) to do so." Although there is disagreement concerning the extension of 'can,' the basic principle is widely shared. Kading (1954), to give an example of classic Kantian scholarship, disputes an 'analytic' interpretation of ought-implies-can, according to which an agent's responsibilities are a function of his current abilities. He points out that even if someone cannot now respond to moral reasons, if the person had an opportunity to develop relevant moral capacities in the past, the person satisfies the Kantian 'can' (which maybe more accurately construed conjunctively as 'can or could have.') Kading gives the example of *feeling appropriate shame*, which a person cannot invoke at will, and yet we can still say, consistent with Kant's intentions, that a person is blameworthy for having neglected opportunities to cultivate appropriate emotional dispositions in the past. Importantly, Kading states that, "it would clearly be absurd to command the impossible" (p. 13), and yet it is reasonable to hold people responsible for having done nothing to prevent undesirable states of affairs. Thus, his view, all things considered, closely resembles my own, and respects the basic principle of avoidabilty.

Pluralistic utilitarian accounts are capable in principle of accepting an avoidabiltiy constraint on responsibility. However, some utilitarian scholars reject heteronomous principles intrinsically. One putative example is Smart (1961), who offers a doctrinaire utilitarian account of moral responsibility. On close scrutiny, however, it is unclear whether he succeeds in defending his intuitions about particular cases on utilitarian grounds alone. Smart asks us to imagine two school boys who fail to do their homework, one of whom is "lazy," and the other of whom is "stupid" (p. 68). To avoid archaic and politically incorrect language, let us suppose that one boy is careless and the other has a learning disability. According to Smart, the first boy deserves blame but the second does not, because only the first boy is capable of responding to recrimination. Although this is a utilitarian explanation, it implicitly invokes the notion of avoidability, by suggesting that one boy is *incapable* of improving, i.e., his failing is unavoidable. So it seems that

avoidability is required to explain why certain actions are sanctioned by utilitarianism.

Moreover, one can imagine cases in which avoidability trumps utility, and we must act against the greater good. For example, if failing students with learning disabilities would enhance the class average, it would not be permissible to pursue this course of action, as it would be unfair to the victims. (Let us permit that assigning a failing grade is a form of negative reactive attitude, as per my preferred wide reading of Strawson.) I do not think that anyone with the least degree of moral integrity would want to bite the bullet on this policy. This objection resembles Williams' (1973) argument against utilitarianism (as a normative theory,) wherein he asks us to imagine a small racial minority that is so hated by the majority, it would maximize utility to annihilate them through genocide. Our sense of moral integrity, Williams says, prevents us from endorsing genocide on principle, even if it is utilitarianly optimal. So utilitarianism must get something wrong. If so, then normative utilitarianism must be misguided; but I believe that utilitarian accounts of *responsibility* are misguided in a distinctive manner, as they neglect the indispensability of avoidabilty. Williams is right that utilitarian accounts of permissibility neglect important values; but utilitarian accounts of responsibility uniformly neglect a distinctive principle—namely, avoidabilty. A second insight that we can glean from Williams is that utilitarianism neglects the importance of the subjective perspective. Even if *one* (abstract agent) should kill one person to save five others, it does not follow that *I myself* should kill one person to save five, since doing so may compromise my moral integrity. Similarly for moral responsibility, we can say even if *one* (abstract agent) is blameworthy for failing to complete the class assignment, it does not follow that *I myself* am blameworthy for this transgression, since I may have a legitimate excuse. In judging responsibility, we must assess a *particular* agent's capacities relative to a general normative principle. This is why avoidabilty is an essential condition of moral responsibility.

We have been discussing educated discretion, and one may worry that SRP requires too much discretion (or practical wisdom, perceptual acuity, affective sensitivity, or what have you) to be applied effectively in particular cases. But this is a mistake. Applying SRP is a straightforward case of inductive reasoning, which is involved in all moral deliberations about general principles. Moral philosophers generally agree that applying general principles to particular cases requires a well-honed sensitivity to salient factors. This is why all ethicists (Kant, Mill, Aristotle) provide some account of moral education, and also agree that no matter how morally educated one is, judging particulars will still be difficult. This is why there has never been consensus amongst moral philosophers: people disagree not only about first principles, but also about how to apply those principles to particular cases, which is one of the reasons why trolley problems will never die. (These dilemmas elicit particular disagreement amongst utilitarians, who are undecided about whether or not it would maximize utility to kill one person in order to save five, but Kantians also

fail to achieve complete consensus.) Likewise, it will inevitably be difficult to apply SRP to concrete cases, but practice and education will help. Further, if the analogy between law and morality holds, there may an equivalent to landmark legal cases in moral responsibility, which can set precedents acknowledged by disputants.

The difficulty of exercising responsible discretion does not count especially against SRP. Indeed, the preceding paragraphs illustrate just how difficult it can be to apply utilitarianism, Kantianism, and virtue ethics to a case as ostensibly simple as adultery—something we encounter very often in ordinary life, if the statistics are accurate (According to the *Kinsey Institute for Research in Sex, Gender, and Reproduction*, 17.2% of cohabitating couples and 25.1% of married couples admit to committing adultery in 2011.)[8] Obviously, one cannot adjudicate the moral status of a particular act of adultery until one has all the relevant information, including contextual and motivational data. Even utilitarianism, which (on most accounts) purports to be particularly simple by virtue of reducing moral deliberation to a matter of quantification, nonetheless requires a difficult computational processes, wherein we estimate the likely long-term consequences of an action (indefinitely into the future), the likely collateral ramifications, as well as, on Mill's view, the *quality* of the resultant pleasure (i.e., whether it is 'higher' or 'lower.') So even the simplest theoretical principle requires significant discretionary expertise. The final verdict will be particularly difficult to obtain in 'penumbral' cases, to borrow a term from Hart, where there is no precedent for judging. We should not be surprised, then, that adjudicating moral responsibility is not easy. The adjudicator must reflectively deliberate in light of all relevant information. In the case of the student who fails to submit her assignment on time, one must consider whether the student has a relevant physical, intellectual, or mental health disability, or personal issues such as death in the family, a medial emergency, parents divorcing, recent trauma, and so on. To simplify this judgment, universities typically codify valid excusing condition, but this does not rule out personal discretion. As an instructor, I have had to consult the undergraduate program director several times to ask if a given excuse was valid—for instance, one student found herself suddenly homeless, another suspected that he had a psychological disorder, but was not yet registered with mental health disability services, as this process takes time, and yet another student was subject to physical abuse by a family member. These are discretionary cases which I had to decide myself. My claim, note, is not that applying SRP is simple (in the sense of not requiring any moral wisdom,) as I do not think any standard of moral responsibility can be applied without a degree of wisdom. My claim is that SRP is the most simple *explanatorily adequate* account of moral responsibility. This is all that is required to satisfy Occam's Razor, since this principle requires explanatory adequacy. Frankfurt's view may be more streamlined, so to speak, since it only takes into account one variable, i.e., the agent's desirative

[8] http://news.nationalpost.com/2012/11/16/graphic-the-demography-of-adultery/

profile, but it faces several difficulties, both methodological and explanatory in nature. Namely, (i) it is impossible to judge a person's desirative profile without considering observable evidence, (b) a person's desiderative profile is inadequate as a basis for moral responsibility, since it excludes situational factors that may cause or condition character flaws, and (c) Frankfurt's account generate counterintuitive conclusions, as we saw in the hypothetical example above, in which a student spent all weekend playing video games but did not reflectively endorse her actions. On the second score, we have seen cases of coercion that cannot be accounted for in psychological terms, such as Ishmael Beah. So my claim, in sum, is that SRP is as simple as one can reasonably wish, provided that one is not willing to cut explanatory corners.

The last concern that I will address here, which was raised to me by Robert Myers, is the worry that moral-SRP ignores important distinctions between legal and moral responsibility, such as the fact that the law sometimes sets higher standards of proof than morality, and that the law sometimes imposes strict (impartial) standards of liability. I am sympathetic to this concern, but I do not believe that legal and moral responsibility come apart at these junctures. Rather, if my account of moral-SRP is right, it follows that moral responsibility can be strict, for the same reasons that legal responsibility can be strict, viz., to distribute responsibility fairly and efficiently. Before defending this justificatory principle, consider some intuitive examples.

(i) First, breach of trust. If the trustee of an estate squanders all of the money in it, then the person is legally responsible to compensate the defrauded beneficiary regardless of *mens rea*. But I do not see why the same judgment would not hold of moral responsibility, which, I have taken pains to show, applies to cases of negligence and recklessness. Surely we do not want to say that the negligent or reckless trustee is not responsible to provide restitution. On the other hand, suppose the trustee commits breach of trust due to severe mental disorder. According to the law, the trustee would nonetheless be responsible to repay the defrauded beneficiary if possible (that is, if she has the funds). But what does *moral* responsibility entail? Do we not also want to say that the trustee is morally responsible to repay the lost funds, even though she was not morally responsible for having lost them in the first place? I think that we do, on intuitive grounds; but furthermore, based on considerations of (a) social wellbeing and (b) fairness to the victim, this conclusion naturally follows. Thus, SRP collapses the distinction between moral and legal responsibility.

(ii) Second, vicarious responsibility. Suppose that a person's unleashed dog bites a small child. The pet owner will be held vicariously legally responsible, even if she did not intend harm. But what of moral responsibility? Intuitively, I believe that we share a

tendency to blame the pet owner; and certainly on SRP, based on considerations of social welfare and fairness to the child and child's family, blame is warranted. A similar case involves a bartender's responsibility to deny alcohol to an inebriated patron. If the bartender ignores the law and the patron subsequently causes harm to someone else, the bartender is legally responsible. Does morality agree? I think so. I believe that we would intuitively hold the bartender morally responsible for any harms caused by the inebriated patron, especially in light of the fact that the law expressly forbids the bartender's culpable action. The law, in other words, provides a moral reason to refrain from serving intoxicated patrons, in light of the considerable social risk that it carries, and the law's authority. Moral-SRP explicitly blames the derelict bartender, on the basis of fairness and social welfare. So once again, legal and moral responsibility converge.

(iii) A more controversial case, at first blush, is trespassing. According to the law, a person may be strictly liable for trespassing, even if she did not realize that she was intruding on private property. There are two ways of assessing this case from the moral perspective. First, it is possible that the intruder could not reasonably have known that she was trespassing, because there were no visible markers, such as a 'No Trespassing' sign. In this case, I believe we would agree, the person is not morally responsible because she is not committing a moral infraction. Rather, the property owner is at fault for not providing proper markers. Indeed, due to the property owner's negligence, the intruder is not legally responsible, either. On the other hand, the intruder may have trespassed due to a failure to mind her surroundings, in which case she is guilty of negligence. On this interpretation, she is both legally and morally responsible, since she should have known better. Once again, legal and moral responsibility coalesce.

These examples are meant to show that my conception of moral responsibility corresponds with legal responsibility in an intuitive way—that is, the conclusions of this inquiry are supposed to match our intuitions fairly closely. On the other hand, it may seem counterintuitive to think that moral and legal responsibility should be isomorphic in the way that I have suggested, especially considering that few philosophers have advanced this position. One notably exception is Peter Cane (2002), who is first and foremost a legal scholar, but argues that legal and moral responsibility are essentially "symbiotic" (p. 163). Nonetheless, he grants that the law is more institutional than morality, because it is interpreted and applied by authoritative tribunals. But he takes this to imply that, *contra* accepted wisdom, the law "can make a contribution to thinking and judgment of responsibility outside the law" (p. 12), bringing the two domains closer together. The

main thesis of his book is that the law's rich detail can provide a lesson to moral philosophers, whose conception of responsibility is eminently vague. (This parallels me own view.) Furthermore, Cane believes that philosophers typically misunderstand the relationship between the law and morality based on an assumption that morality is "a source of values whereas law is purely conventional" (p. 12). Hart, for instance, says that morality is "objective," whereas the law is constructed out of interpersonal agreement, so morality is, in a sense, ontologically deeper, and rests on a separate justificatory basis. This supports the view that morality and law are "totally separate normative domains that exist side-by-side but do not interact with one another" (p. 13). Cane believes that this construal is both misguided and unhelpful, and proposes that, "to the extent that legal rules and principles of responsibility coincide… with morality, they could be viewed as a reflection and reinforcement of it" (p. 13). On one side, the law can concretize commonly accepted moral principles, and on the other side, laws can create moral obligations by generating contracts, promises, and obligations. This is not entirely to the point, however, since Cane is describing the relationship between moral and legal permissibility, as opposed to responsibility *per se*. But it is instructive that even this broader relationship may be much closer than people presume.

Regarding responsibility, Cane argues that "social values" are an "important focus of legal responsibility practices which typically finds little or no (explicit) place in many philosophical [moral] analyses of responsibility" (p. 53). His complaint is that philosophers tend to conceive of moral responsibility in atomistic terms, neglecting the important social dimension of this practice. For this very reason, they tend to see the law and morality as differentiated by the former discipline's 'conventional' nature; and yet on close scrutiny, *both* practices are inherently social and, in this regard at least, conventional. Thus, the posited distinction between the two disciplines is largely illusory, generated by a neglect of social relationships. Furthermore, moral responsibility, like legal responsibility, cannot be properly understood without attending to the *distributive* responsibilities generated by the interpersonal relationships that characterize moral life. It is worth citing Cane's argument to this effect at some length:

> Because responsibility practices rest on general principles according to which responsibility is allocated and distributed, holding any particular individual responsible in any particular set of circumstances has potential ramifications for other individuals who find themselves in similar circumstances. The point can be simply illustrated by referring to the famous case of Donoghue v. Stevenson, in which it was decided for the first time in England that a manufacturer could be held liable to a consumer who suffers injury as a result of a defect in a product caused by the negligence of the manufacturer. At one level, the case concerned the responsibility of a particular manufacturer to a particular consumer. At another level, it dealt with much larger social questions about the *relationship between manufacturers and*

> *consumers generally*. In other words, the case was not only about the responsibility of one individual to another, but also about *the distribution of rights and obligations in society generally*. For this reason, in deciding how to resolve the case, the judges considered not only the issue of "fairness" as between the plaintiff and the defendant, but also the wider social and economic impact of a decision one way or another in relation to the two parties before it. (p. 53, emphasis mine)

The key point for our purposes that moral responsibility is not an agent-focused concept, but rather a distributive concept, consisting in practices and principles that enable us to distribute responsibility fairly and effectively. The case of Donoghue v. Stevenson, above, illustrates that even if the manufacturer did not intend to cause harm, he owes reparations to the injured party. Thus, the criterion of *mens rea* is discounted, and responsibility emerges from the nature of the fiduciary relationship. This verdict is justified, according to Cane, because it reflects a distributive principle that achieves (a) fairness to the plaintiff in light of the damages incurred, and (b) indemnity from undue social harms stemming from corporate negligence. This, in effect, echoes my account of moral-SRP, which aims to distribute responsibility appropriately on the basis of fairness and social welfare, optimally balanced in reflective equilibrium. The correspondence is coincidental, but Cane' articulation of the argument may help to shed light on my own rationale.

Cane's work also sheds light on strict responsibility in law and morality. One might assume that *strict* responsibility does not involve avoidabiilty, since it ignores *mens rea*. For instance, in breach of trust, the trustee is responsible to compensate the beneficiary even if he did not intent to cause harm; so it seems that the trustee could not have done otherwise. Yet Cane debunks this myth, arguing that while legal responsibility does not require *fault* (in the technical legal sense,) it *does* assume the ability to do otherwise.

"Fault," Cane explains, denotes "conduct that breached a legally specified standard of conduct… regardless of whether the conduct was accompanied by any particular mental state" (p. 82). So an agent can be responsible in the absence of 'fault.' But from this, it does not follow that the agent *could not have* done otherwise. For example, from the fact that the trustee did not subjectively intend to cause harm, it does not follow that he could not have been more vigilant, more informed, or more sensitive to the risks he was taking. Cane describes the typical mistaken inference from lack of fault to lack of ability to do otherwise as follows:

> Having made the assumption that strict liability is liability in the absence of fault, the conclusion is then (often implicitly) reached that there is nothing that the person held strictly liable could have done to avoid the liability—or, in other words, that the events that gave rise to the liability were (in some relevant respect) outside their control and, therefore, a matter of bad luck. Unfortunately, this conclusion is false even allowing the premise. Of

> course, if the person held strictly liable was actually at fault, the liability-attracting events may have been relevantly under their control. But even if they were not at fault, it does not follow that the occurrence of the liability-attracting event was beyond their control… It does not follow from the premise that an event was not a person's fault that they could have done nothing to prevent it. (p. 84)

Consequently, it would be wrong to think that strict legal responsibility does not have an analogue in ethics. According to moral-SRP, we ought to assign 'strict responsibility' in cases of negligence and breach of trust, just as we do in legal cases. It would be odd, I believe, to say that the corporate owner who failed to supervise his recalcitrant employee is not morally responsible for negligence, and odder still to say that he is not responsible to compensate the victim, which is itself a form of responsibility on the interpersonal model.

Now one might think that moral responsibility has no analogue for tort law, which is a civil wrong that causes someone to suffer a loss, for which the tortfeasor is legally responsible and required to pay damages. While it is true that Strawson's reactive attitudes do not explicitly capture this type of response, this does not mean that they exclude them. Strawson's account, I believe, is far too vague to capture the full gamut of appropriate responses to various infractions. This is one of the reasons why I agree with Cane that legal scholarship can help to refine and educate our notion of moral responsibility, which lacks fine-grained texture. It is perfectly reasonable and intuitive, I believe, to countenance the imposition of compensation, restitution, or reparations as a form of reactive attitude, and thus a form of moral responsibility. In conversational English, we are wont to say that when a trustee has defrauded someone of funds, the person is *morally responsible* to pay restitution. Again, this is partly because trustees are responsible, due to their professional office and consequent relationship to their clients, to protect their clients' interests. Our judgment of responsibility to pay compensation rests on an acknowledgement of the moral status of this relationship as generating responsibilities. Moral-SRP recognizes this relationship as morally valuable from the perspective of fairness and social welfare, and from the distributive nature of responsibility burdens. Thus, although Strawson does not explicitly mention requirements of compensation, we might incorporate such actions under the rubric of the moral reactive attitudes without harming the theory. Indeed, doing so may be necessary to preserve the intuitive and interpersonal aspect of moral responsiblilty, which Strawson greatly valued.

In spite of these arguments, there is a pervasive tendency in moral philosophy to assume that legal responsibility is stricter than moral responsibility, presumably because the moral reactive attitudes are relatively inconsequential. For example, Feinberg (1970) states, "we [moralists] can afford to have stricter standards of culpability than the lawyers since no formal

punishment will follow as a result of *our* verdicts and we do not have to worry about procedural complexities" (pp. 245–246). I would dispute this view, however, because morality, seen as a social practice, can impose severe penalties on those deemed blameworthy. For instance, while it is true that adultery is not illegal, it can elicit severe social sanctions in particular social climates. Ingrid Bergman was temporary exiled from Hollywood, and Bill Clinton's infidelity arguably resulted in personal and professional losses. In extreme cases, people can be subjected to social isolation—complete exclusion form interpersonal relationships—which is comparable to solitary confinement, i.e., the worse form of imprisonment. It would be wrong, therefore, to say that legal penalties are categorically more severe than moral ones. It is more accurate, I believe, to say that the severity of sanctions in either domain should be proportional to the infraction. Furthermore, the burden of proof should be proportional to the severity of the infraction, in both legal and moral responsibility. But this applies to both conceptions of responsibility. So if there is any difference between them, it is one of degree, not kind.

In the next section, I explain how an externalist account of moral responsibility applies to mental disorders.

## Chapter 5: Moral Responsibility and Mental Disorders

### 1. Preliminaries

As an externalist account, SRP might seem unsuitable for determining responsibility for actions stemming from psychological factors such as mental health disabilities.[9] However, this is not so. In this chapter, I argue that, contrary to first appearances, SRP is ideal for adjudicating such cases. This is because the main alternatives, i.e., the deep-self view and the sane deep-self view (as Wolf calls them,) are susceptible to the problems of poor introspective access and poor(er) second-personal ascertainment, as revealed by situationism and cognitive science. This is just as much a problem for psychiatric settings as elsewhere. In addition, I will argue that insofar as internalist theories, as per the traditional conception, deliver zero-sum judgments about responsibility (i.e., either responsible or not responsible,) they tend to reduce persons with mental health disabilities to non-persons, ignoring the range of capacities that contribute to personhood. Moral-SRP, in contrast, isolates and excuses *circumscribed* cognitive deficits, leaving the person's broader cognitive architecture intact. This allows us to recognize that while a mentally ill person may be deficient in one capacity, she may be functional in other morally-relevant regards.

[9] I use the term 'mental health disability,' as per common usage in philosophy of psychiatry, to denote categories in the Diagnostic and Statistic Manual 5 (DSM 5), which includes psychopathy and Tourette syndrome.

This in turn helps to deflate traditional objectifying attitudes toward psychiatric patients, which has plagued psychiatry since its inception.

In this chapter, I Will Cartwright's recent philosophical analysis (2006) as a point of departure for addressing moral responsibility and mental illness. Cartwright considers the two dominant theories of responsibility, i.e., the *reflective-self view* (which is his name for the deep-self view) and the *reasons view* (which is his name for the sane deep-self view,) and contends that neither theory fits with our intuitions about a specific kind of mental disorder, sociopathic personality disorder. He uses the example of the infamous homicidal bank robber Robert Harris to illustrate that we have mixed feelings about mental disorders, but neither of the two candidate theories can accommodate these feelings; and even more problematically, on close scrutiny, both theories "rest on or foster what are arguably fictions" (p. 153). After raising these skeptical doubts, Cartwright surprisingly does not offer a substantive proposal. This leaves us at an impasse, with no positive account of responsibility. This is unacceptable, since we need responsibility to govern and justify our moral attitudes toward others, our business transactions, legal decisions, academic honest hearings, and so on and so forth. To solve this problem without denying the force of Cartwright's objections, I propose an externalist account of moral responsibility that does not invoke the 'fictitious' assumptions of the internalist theories: namely, SRP. I argue that, properly understood, SRP matches our intuitions about mental illness, and avoids the drawbacks of the disputed alternatives.

In my penultimate section, I apply SRP to the case of Tourette syndrome, to show that this standard provides a better theoretical basis than internalist accounts for our broadly-shared intuitions about Tourettic individuals: in particular, the intuition that these individuals are not morally responsible for their tics, or the socially inappropriate behaviour that often results. This runs counter to Shroeder's well-known deep-self account of Tourette syndrome (2005), which, I argue, rests on a faulty moral psychology, one that attempts to clearly demarcate desires from non-desires, and to assign responsibility only to the former. Even if it were possible to make this bifurcation, which I doubt, I do not think that it gives us the results we want.

In the final section, I address objections raised to me in correspondence by Jeffrey D. Bedrick, clinical psychiatrist at Dexel University College, and Christian Perring, professor of philosophy at Dowling College. This response will help to resolve outstanding issues that were not addressed in the original manuscript.

**2. Introduction**

It is contested whether and to what extent moral responsibility can be ascribed to persons with mental health disabilities. Cartwright evaluates two prevalent theories of responsibility in terms of their suitability for morally appraising sociopathic personality disorder, particularly as embodied in

the example of the famous homicidal bank robber, Robert Harris. Cartwright argues that our intuitions about Harris conflict because we are viscerally horrified by Harris' actions, but we are forced to reconsider our initial reaction when we reflect on his history of extreme abuse and neglect. After thoroughly considering two 'candidate theories'—the *reflective-self view* and the *reasons view*—Cartwright concludes that neither is able to resolve this dilemma, partly because both "rest on or foster what are arguably fictions" (p. 153). This analysis leaves something to be desired—namely, a positive theory of moral responsibility that can be applied to cases of sociopathic personality disorder, and appealed to to explain or resolve our conflicting intuitions.

In his response to Cartwright, Benedict Smith (2006) notes this lacuna. He writes that Cartwright sets out to "illuminate both the theory and practice of holding people responsible" (p. 144), but ultimately illuminates nothing more than that, "the practice of responsibility depends on one sort of pretense or another" (p. 155). The problem with Cartwright's analysis is that it gives us no principled means of differentiating between culpable wrongdoing, and non-culpable wrongdoing due to mental illness. It provides no general criteria, for instance, for differentiating between a fully responsible killer, and a temporarily insane one; a new mother who culpably neglects her child, and one who does so only because she is suffering from post-partum depression; a deliberately rude person, and a person who makes socially inappropriate comments due to Tourette syndrome, and so on. Doing without moral responsibility is not an option, since this concept plays a vital role in structuring our social relationships and maintaining social order. (To give but one example, how shall we distribute reasonable accommodations to students with disabilities if moral responsibility is a fictitious concept?) But Cartwright provides us with no viable alternative. As Smith notes, if Cartwright wishes for us to reject fictitious theories, "then we might plausibly expect some replacement. But what would a 'factual' or 'real' theory of responsibility look like?" (p. 166)

In this chapter, I aim to answer to this question. I offer a third candidate theory of moral responsibility, which, I believe, avoids the problems that afflict the other two views. For the purpose of pursuing this analysis, I will assume that there are objective moral facts—for instance, that it is objectively wrong to murder an innocent person for no reason. This undercuts Cartwright's objection that moral realism may "be revealed as a convenient fiction" (p. 154). However, I take this to be a very modest and reasonable assumption, which is prerequisite to developing any positive account of moral responsibility, and it is furthermore borne out by our practice of holding each other responsible, which is humanly unavoidable. As Strawson puts it (1963), a sustained suspension of moral appraisal, "and the isolation which that would entail, does not seem to be something of which human beings would be capable, even if some general truth were a theoretical ground for it" (p. 81). Each of us *in practice* necessarily endorses the reality of moral responsibility many times over in the course of every day, so to deny its

theoretical existence would amount to hypocrisy, a violation of the rational principles that ground philosophical inquiry.

What I take to be a more vexing problem is what Cartwright describes as "perhaps the most important traditional objection to responsibility," i.e., the idea that responsibility "rests on the fiction that in most situations we can do otherwise than we do" (p. 153). My account may seem *prima facie* to succumb to this problem, but only if we assume a traditional, metaphysical interpretation of 'can do otherwise.' In philosophy there are several well-established theories of how to interpret this phrase, and I defend one based on Miranda Fricker's feminist epistemologist account of responsibility (2007). This interpretation provides a practical and persuasive method for establishing when it reasonable to hold an agent responsible for wrongdoing in light of her capacities and circumstances, and thereby captures a range of intuitive excusing conditions. This in turns provides the foundation for a realist theory of responsibility which allows us to discriminate between culpable wrongdoing on the one hand, and non-culpable wrongdoing on the other, which is excused on the basis of an agent's observable cognitive deficits and epistemic environment.

**3. The Reflective-Self View**

Before assessing the reasons view, we must consider Cartwright's first candidate theory, the *reflective-self view*, which has been dominant throughout most of the last century. It entails, in Cartwright's words, that "for a person to be responsible for an action requires, not merely that it flow from some desire of the real self, but that it flow from that desire of the real self that the reflective self has determined should issue in action at that moment" (p. 145). The same view has been critiqued somewhat more comprehensively by Susan Wolf (1987), under the name of *the deep-self view*, which she attributes in its exemplary forms to Frankfurt, Watson, and Taylor (p. 375). Wolf says that this view has two merits: (i) it allows us to differentiate between our normal selves and cases of kleptomania, brainwashing and posthypnotic suggestion (in which the agent's deep self is severed from her will and she is not responsible); and (ii) it avoids the problem of metaphysical determinism, according to which everything that we do is predetermined by an infinite chain of antecedent causes, making it impossible for us ever to be responsible for anything.

However, this view also has two crucial flaws, which are emphasized by both Wolf and Cartwright. First, it gives rise to "an apparently indefinite series of levels of reflection of progressively deeper selves" (Cartwright, p. 147), making it psychologically impossible to achieve the level of self-scrutiny required for responsibility; and, even if we accept a lesser degree of self-reflection, "the positive endorsement of character by reflective selves would not seem to be frequent. To pretend otherwise, which it seems the theory must do to explain how most people

are responsible, is to embrace a fiction" (Cartwright, p. 154). In other words, this view inadvertently makes responsibility a psychological impossibility.

The second main objection is that the reflective-self view does not give an adequate account of how to ascribe responsibility to people with sociopathic personality disorder, and other mental health problems. To see this, consider Harris' personal history in more detail. On Cartwright's description, Harris

> was born two and a half months prematurely when his mother was kicked in the stomach by her husband, a violently jealous man who thought that the child was not his. Both parents were alcoholic. The father frequently beat his children, often causing serious injury, and he was convicted of sexually molesting his daughters. The mother was arrested several times, once for bank robbery…[Harris suffered] from a learning disability, he was teased at school and felt stupid… At fourteen he was sentenced to a youth detention center for car theft. While detained he was raped several times and he twice attempted suicide. (p. 144)

When we compare Harris' homicidal actions against his pitiful personal history, we tend to experience mixed emotions: "The feelings of outrage and loathing inspired by the former collide with the feelings and thoughts induced by the latter, which may be rather various" (Cartwright, p. 144). Cartwright does not immediately try to resolve this dilemma. But he later provides a tentative solution: "*if* the truth about Harris was that he was simply a man who had desires and acted on them, who cared not at all about what his desires were, an amoralist or a sociopath perhaps, then on the reflective self view he was not responsible because he *had no reflective self*" (p. 148, emphasis mine.) This raises some interesting questions. For instance, how are we to know if the antecedent of this conditional is true? Is it possible for a person to have *no* reflective self? And what about people who do have a reflective self—are they to be held responsible for *everything* they do regardless of the circumstances? These questions provide a basis for three objections to the deep-self view:

(i) Since the reflective-self view is internalistic (i.e., based on the agent's inner states), it may be impossible to discern whether or not an agent has a reflective self. Cartwright himself cannot decide whether Harris has one, since Harris' description of his past suggests that he "seems to be accepting responsibility for his life generally, including presumably the murders," but on the other hand, "even to say that Harris is endorsing, or just standing by, his actions may itself be to read too much into the remarks" (p. 148). Since Cartwright cannot decide one way or the other, he asks us to "suppose… for the sake of argument, that Harris could be construed as having

endorsed the desires on which he acted" (p. 148). Then Harris would seem to be responsible for his actions. But, Cartwright continues, even if Harris did endorse his homicidal desires, "the evaluative criteria revealed by such an endorsement will be so grotesquely distorted as at least to put in question Harris's responsibility" (p. 148). So we are left with further questions, *ad infinitum*. These hypothetical claims and unanswered questions cast doubt on the practical utility of the reflective-self view, which, though it is a nice theory, does not seem capable of resolving our questions about when a person is genuinely responsible. In real life these questions cannot simply be left open, since responsibility is a central mechanism in dispensing criminal justice, organizing social life, and maintaining relationships.

(ii) On the reflective-self view, a person cannot be held responsible if she has no reflective self. But is it reasonable to say of any person that she has *no* reflective self? To say that someone has no reflective self is to say that she is not a moral agent—that is, in philosophical terms, not a *person.* A person is not merely a genetic human, but a creature with moral qualities such as rationality, emotionality, and a capacity for self-reflection (Warren 1996, p. 130). A non-reflective human can be likened to what Strawson calls a "warped or deranged" individual: "an object of social policy…, a subject of what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided" (Strawson, p. 79). Although Strawson intends for this attitude to include a variety of emotional responses, even on a liberal interpretation, the language of objectification has discomfiting implications, especially when we consider the egregious history of discrimination against and mistreatment of people with mental disorders. As C. Perring (2009) notes, "We do not want to take a completely objectivizing stance towards people with schizophrenia [for example], as Kennett (2007) points out in connection with Peter Strawson's ([1962] 1982) distinction between participant reactive attitudes and objective attitudes, because this reduced the mentally ill people to objects with no human aspect, and takes them out of the community of other humans" (p. 2). In a similar same vein, H. Pickard (2011) observes that, "compassion and empathy are central to good therapeutic care (Gilbert 2010). They are essential when working with service users with PD [personality disorder]. The reason is simple: a compassionate, empathetic stance is at odds with a blaming stance" (p. 220). Pickard continues, "one central way that clinicians can achieve compassion and empathy toward service users is simple: proper attention to service users' past history" (p. 220). These perspectives suggest that the reflective-self view does not

provide a sufficiently nuanced view of moral responsibility—one that would permit us to excuse a person from responsibility for a particular action or motive, while still viewing the agent as a person, and *ipso facto* a subject of compassion and empathy.

It is also significant that the 'non-person' account conflicts with many people's subjective experience of having a psychological disorder. In this connection, the autism rights activist Jim Sinclair (1992) writes,

> "my personhood is intact. My selfhood is undamaged. I find great value and meaning in my life and I have no wish to be cured of being myself.... Grant me the dignity of meeting me on my own terms—recognize that we are equally alien to each other, that my ways of being are not merely damaged versions of yours" (p. 302).

Sinclair's affirmation of his intact personhood is corroborated by his distinguished body of work on autism, which testifies to his possession of (at least) rationality, emotionality, and self-awareness. And yet autism is a clear-cut case of "psychological abnormality" which would exclude Sinclair from personhood on the standard view of moral responsibility advanced by Strawson. Yet there is no self-evidently good reason to eliminate such persons from moral society.

(iii) A third problem with the reflective-self view is that it does not seem to allow us to excuse people who have reflective selves but are unavoidably ignorant. JoJo is an example. He is, by all appearances, congenitally cognitively normal, but cannot grasp or appreciate moral facts due to his social position. Greenleaf is another example. Notably, Fricker mentions the distinction "between exceptional and routine moral judgment," which "points to the possibility of a more nuanced range of moral attitudes to historically and culturally distant others—for it allows us to avoid the hubris of deeming them blameworthy for actions not routinely regarded as wrong in their culture, while still holding them morally responsible to this or that extent, depending on how nearly available the exceptional moral move is judged to have been" (p. 105). This reference to a 'nuanced' range of attitudes highlights the need for a *pro tanto* account of responsibility, which holds a person responsible to a certain degree, depending on the facts of the case. A person from a less sexist environment might be excusable to a lesser degree, or not at all, depending on the individual's particular social location.

Although Wolf and Fricker provide merely anecdotal accounts, their examples can be explicated in terms of deeper theoretical principles. Specifically, they can be

interpreted in terms of SRP, which requires a responsible agent to have, in Hart's words, "the normal capacities, physical and mental, for doing what the law requires and abstaining from what it forbids, and a fair opportunity to exercise those capacities" (Hart, p. 152). On this view, *avoidability,* either psychological or circumstantial, is the central evaluative criterion. This approach differs from the reflective self view in that it implies that even if a person has a fully reflective self, she can still be excused from responsibility provided that she lacks access to certain epistemic goods due to her circumstances—for instance, if acquiring a certain prejudice was virtually unavoidable in her ethical culture. Since liability and responsibility both depend heavily on the notion of avoidability (a point that I will elaborate on in the next section), fully self-reflective agents can be excused of responsibility on these epistemic grounds.

Given these problems with the reflective-self view, it is worth considering a second account of responsibility.

**4. The Reasons View**

The reasons view adds an additional condition to the reflective-self view. It holds that "responsibility requires more than just control; it requires both cognitive and moral understanding" (Cartwright, p. 149). This view is defended by Wolf under the name, *the sane deep-self view*, which requires that an agent be 'sane' in the relevant sense.

This view has three merits. First, like DSV, it allows us to differentiate between our normal selves and cases of kleptomania, brainwashing and posthypnotic suggestion. Unlike DSV, however, it avoids the infinite regress of reflective selves that besets the reflective-self view, since it grounds the agent's freedom not merely in volitional endorsement, but in objective facts. That is, the agent must not only endorse his desires, but see the world for what it is. Thus, infinite reflection is undercut by a prior, foundationalist requirement. And third, as Wolf points out, it "deals with the case of JoJo [the son of an evil and sadistic dictator] and related cases of deprived childhood victims in ways that match our pretheoretical intuitions" (p. 382).

In spite of these merits, the reasons view faces difficulties that neither Cartwright nor Wolf seems to realize. Specifically, it is vulnerable to the second set of difficulties that besets the reflective-self view. To whit:

(i) Since the reasons view is internalist (i.e., based on the agent's inner states), it may make it impossible to determine whether the agent is either blameworthy or 'insane.' It is clear that Harris' homicidal actions reflected hopelessly perverted desires, but

does it follow that he lacked cognitive or moral understanding? Perhaps he could have exercised control over his volitions at some point in his personal development but chose not to. That is, perhaps he was culpably negligent. Cartwright gives us rough guidelines for how to resolve these questions: “In cases where other factors such as fear and fatigue are absent and the moral reasons are at their strongest, the failure to do the right thing invites the explanation of incapacity” (p. 151). But this only pushes the question back one step: How do we know if factors such as fear and fatigue are present? And even if moral reasons are pervasive, how do we know if the agent is paying due attention to them? These are the questions that arise when we inquire into an agent’s subjective mind.

Solving this problem, I submit, requires an objective account of responsibility, one that defines responsibility in terms of factors external to the agent’s mental states. This is what SRP does, and what it was intended for. It measures responsibility in terms of what would be blameworthy in the average person, ‘the person of ordinary intelligence and prudence.’ In the next section, I will explain how this account can be used to solve the problem of Harris’ moral status.

(ii) The second problem that the reasons view shares with the reflective-self view is that it tends to depersonalize the subject and alienate her from basic empathy and compassion. If the individual is ‘insane,’ then she is not one of *us*—a member of the moral community—but rather, “an object of social policy,” a subject of “treatment…, to be managed or handled or cured or trained; perhaps simply to be avoided” (Strawson, p. 79). However, since a pathological moral deficit (which is beyond the agent’s volitional control) is a specific type of *cognitive deficit*, it is a circumscribed deficit which leaves much of the moral agent intact. Psychopathy, for instance, is thought to be due either to an empathy deficit (Nichols 2004), or an informational processing deficit in frontal lobe systems (Murphy 2006), but in either case, psychopaths may maintain other aspects of personhood such as agency, intentionality, a capacity for self-reflection, and, depending on the case, a significant degree of rationality and emotionality. Moreover, even if a cognitive deficit is pathological, it may be repairable through cognitive behavioural psychotherapy or other interventions, but only if the agent is first treated as a person capable (to some extent) of social understanding and critical self-reflection.

(iii) The third difficulty is that the reasons view seemingly does not allow us to excuse fully sane people who make moral mistakes due to unavoidable ignorance. This is demonstrated in Fricker’s example of Greenleaf, who is described as a victim of

> "circumstantial epistemic bad luck" (p. 33). I believe that SRP is better equipped to deal with these kinds of cases, since it defines responsibility in terms of what an agent could reasonable be expected to do based on both the agent's capacities and the *circumstances* in which these capacities were formed and developed. It does not restrict excuses to the realm of *endogenous* cognitive/moral deficits. Owing to this condition, SRP allows us to acknowledge culpable negligence, in which the agent's subjective intentions are irrelevant, since she ought to have known better in light of endogenous considerations (regarding her circumstances and personal history.) More on this later.

These kinds of concerns prompt us to consider a third account of responsibility. In the next section, I argue that SRP retains all of the merits of the reasons view, but avoids the problems that afflict that account.

**5. SRP and Insanity**

SRP, as we know, is an influential account of liability in criminal law and jurisprudence, but it is seldom discussed as a candidate for determining moral responsibility. On Hart's conception, a responsible agent must have both the capacity required to comply with the relevant norms and a fair opportunity to exercise those capacities, since avoidability—the capacity to avoid violating relevant norms—seems central to responsibility (Hart, p. 154). By contrast, Wolf's view emphasizes 'sanity' rather than avoidability. But it is unclear from her JoJo example that *sanity* is doing the real work. On close scrutiny, what seems crucial is not that JoJo is *insane*, since he could conceivably have become insane through culpable negligence—for instance, if he had willfully ignored, at every turn, the social signifiers in his culture that indicate that torture is wrong. Rather, what seems to be pulling at our moral intuitions is the fact that JoJo grew up in 'a small, undeveloped country' in which, it is reasonable to assume, alternative viewpoints were not reasonably available. If the example were transposed to modern-day America, JoJo's actions would not be deemed excusable, since America is a democratic, technological, highly educated society, where informational resources are ubiquitous. A sociopathic person in such a context may very well be responsible for his moral insanity. It is this expectation that justifies the legal principle of *ignorantia juris non excusat* (ignorance of the law does not excuse): insofar as the American legal system is committed to promulgating the law in terms that are intelligible to ordinary Americans, and providing avenues for appeal and reform, it is justified in holding ordinary citizens responsible for acting lawfully. The same principle, by parity of reasoning, applies to moral responsibility: it is fair to hold an ordinary person responsible for a moral

transgression, unless that person could not have done otherwise due to radical epistemic constraints.

This view has all of the strengths of the previous two theories, but none of the weaknesses. First, it allows us to differentiate between culpable wrongdoing and cases of kleptomania, brainwashing and posthypnotic suggestion, in which the agent could not have done otherwise due to volitional or cognitive impairment. Secondly, it avoids the infinite regress of reflective selves that besets the reflecive-self view, since it does not depend upon self-reflection. And thirdly, it matches our pretheoretical intuitions about JoJo and related cases of deprived childhood victims, and provides a justificatory basis that is more plausible than Wolf's. That is, avoidablity is a more compelling explanation for our intuitions than sanity, since it is more explanatorily robust. This is clear if we consider cases of insanity brought about by sheer neglect of information, which could plausibly trigger culpability. Only avoidability can account for such culpable negligence.

In addition, unlike Wolf's view and the reasons view, it avoids the three problems outlined above:

(i) First, as an *objective* theory, SRP gives us clearer guidelines than subjective theories for determining when an excuse is present. An external theory is one that relies on observable criteria to measure responsibility—in this case, what can reasonably be expected of an ordinary person under certain circumstances. This standard is more certain than alternatives in that it dispenses with subjective evaluations of the subject's mental profile. It does not need to confirm, for instance, whether an agent's action is the expression of her will, which in turn is an expression of character that is affirmed by her, or if her will reflects an inability to understand and appreciate the world. Instead, it considers whether the agent had access to tools of social interpretation that could have permitted her to act otherwise or develop differently than she did. It may also make use of objective scientific data into the person's cognitive architecture, but this data is quantifiable: for example, the Hare psychopathy checklist.

(ii) Secondly, this view does not depersonalize the target individual because it does not define the subject as either entirely lacking in reflective selfhood or 'insane.' Rather, it isolates excusing factors such as circumscribed cognitive deficits or environmental obstacles (such as gaps in the shared interpretive resources of a society), and defines the agent as constrained in these limited respects, but otherwise normal. For instance, someone with sociopathic personality disorder is a *person* with either an empathy deficit or an informational processing problem in frontal lobe systems; someone with Tourette syndrome is a *person* with a defect of the thalamus, basal

ganglia and/or frontal cortex; someone who acts violently due to severe childhood deprivation and abuse is a *person* with deficient empathy and moral understanding due to her formative circumstances, and so on. These people retain many aspects of personhood—rationality, emotionality, self-awareness, sociability (depending on the case)—even though they suffer from localized deficits. According to influential thinkers throughout the history of philosophy (e.g., Kant, Anderson 1993), features of personhood are the source of human dignity and intrinsic value, as well as what separate human beings from animals or brutes. Thus, it is important not to deprive people of these qualities unnecessarily or on purely conceptual grounds. SRP achieves this goal by focusing on specific deficits, as opposed to global capacities.

(iii) Third, this view allows us to excuse people who are fully self-reflective and 'sane,' but make mistakes due to a lack of epistemological resources, i.e., "a gap in collective hermeneutical resources," in Fricker's words (p. 6). Internalist views, by contrast, do not recognize the force of situational excuses, or, if they do, they provide no philosophical grounds for making this identification. At best, they require an externalist *ad hoc* condition to explain all relevant factors.

Because SRP manages to avoid these three objections, it is an attractive alternative to the reflective-self view and the reasons view. Further, it answers to Smith's request for a 'factual' or 'real' theory of responsibility.

In the next section, I will apply this theory to the case of Tourette syndrome, and in so doing, challenge the prevailing theory of responsibility for neuropsychiatric disorders in philosophy. First, however, it is necessary to resolve the question of Harris' responsibility according to SRP. This will be a somewhat cursory account, which will gain additional illumination in the next section on Tourette syndrome.

On Cartwright's description, Harris's case contains a number of apparently mitigating circumstances. In particular, Harris was severely abused and neglected by his parents, teased at school, raped in juvenile detention, suffered from a learning disability, and likely suffered from morbid clinical depression, as exemplified in his two suicide attempts. Cartwright is ambivalent about whether Harris also had sociopathic personality disorder. In general, Cartwright's description of Harris is vague and speculative, but it strongly suggests a positive diagnosis. Furthermore, in real-life cases, clinical tests could be used to evaluate whether and to what extent someone in Harris' position fits agreed-upon diagnostic criteria. This practical issue, however, is something that all theories must wrestle with. The critical point is that once every possible measure has been taken to assess the subject's condition, the reflective-self view and the

reasons view still fail to take into account all of the relevant data—particularly the epistemic variables.

The problems with the first two views, to summarize is that they rely entirely on the agent's subjective self-reports, which may be illusory or distorted by cognitive bias and the like. (This applies not only to psychiatric patients, but all people.) As Doris puts it, "we can't simply cross-examine [a person] to establish the presence or absence of [extenuating conditions], given the well-known schisms between what we say we care about and what we do; testimony about one's self is very often, intentionally or not, unreliable testimony" (p. 525). From a subjective perspective, one may not experience one's cognitive architecture as affected by an empathy deficit or an informational processing problem in frontal lobe systems, yet one may still have psychopathy. So more than phenomenal assessment is needed to judge responsibility. Secondly, the reflective-self view and reasons view exclude external extenuating conditions, such as cultural isolation. And thirdly, these views implicitly assume a zero-sum conception of personhood according to which the subject is either a full-fledged moral agent, or a 'warped and deranged… object of social policy.' This conflicts with the fact that psychological disorders tend to be circumscribed deficits that leave much of the agent's cognitive capacities intact, as anyone with experience with psychiatric patients can easily attest. Moreover, this explains why many people respond positively to various forms of treatment. By contrast with these views, SRP, as I have explained it, gives due weight to external extenuating factors such as duress, severe childhood abuse, and epistemic isolation. This helps us to resolve the issue of Harris' culpability, based on Cartwright's description. Specifically, if Cartwright's biographical description is correct, then we can infer, at the very least, that Harris' responsibility was significantly curtailed.

**6. SRP and Tourette Syndrome**

There may still be some doubt as to whether an externalist theory is the best method for evaluating cases of cognitive deficits and psychological disorders which are *essentially internal*. There is indeed some difficulty with respect to such cases, but I have argued that this difficulty is generalizable to all theories, as it is inherent to the problem of epistemic access. Nonetheless, to assuage skepticism it may be instructive to compare this view against a reasons view of cognitive dysfunction which has gained considerable credence in analytic philosophy: namely, Timothy Shroeder's theory in *Moral Responsibility and Tourette Syndrome* (2007).

Tourette syndrome is a disorder in which the subject experiences frequent tics consisting of eye blinking, coughing, throat clearing, sniffing, facial movements and coprolalia (i.e., exclamations of obscenities or inappropriate or derogatory remarks.) Shroeder provides a modified reasons account of the moral status of Tourettic tics. He points out that Tourette syndrome is not easily accommodated by the standard reasons view, since

> many Tourettic individuals report that they autonomously, actively choose to tic in many circumstances. They find that not ticcing creates a buildup of tension, as the urge to shout, cough, touch, jerk, etc. does not typically go away but rather intensifies, and so they rationally choose to tic rather than to endure the unpleasant tension created by suppressing the tic. (p. 107)

Furthermore, corpolalia tends to be targeted, in that Tourettic individuals do not usually articulate random insults, but launch insults tailored to their target's specific traits. For instance, they may yell 'fatso' to an obese person, or 'klutz' to a clumsy person (Shroeder, p. 114):

> a Tourettic individual who tics often takes an action which reflects his considered judgment, which is in line with his values, which has higher-order conative endorsement, which fits with his long-term plans, which responds to reasons he has, which he could have chosen not to take (Shroeder, p. 108).

This implies that Tourettic individuals meet the criteria for responsibility on the reasons view: they understand the world and act autonomously, and their tics express their considered judgments and evaluative beliefs. (e.g., They believe that a person is obese, and shout 'fatso.') Thus, on the standard reasons view, they must be deemed responsible for their tics. However, Shroeder is "very reluctant" (p. 115) to accept this implication, because there seems to be a marked difference between Tourettic individuals and non-Tourettic individuals at the level of motivation. Therefore, he tries to salvage the reasons view by eschewing "the Smith-style motivational theory of desire" (p. 115), which depicts all motives as desires. More specifically, he tries to use cognitive neurobiology to argue that tics cannot be 'desires,' because they stem primarily from the basal ganglia, which "produces the movements learned through operant conditioning" (p. 117). Consequently, if tics are desires, then so are all operantly conditioned behaviours, including, say, the tendency of a driver to turn left at a familiar intersection by dint of habit. But this cannot be a desire, so Tourettic individuals cannot be responsible for their tics. So the argument goes.

The main problem with this analysis, I submit, is that even if Tourettic tics are not desires (which is quite contestable given that they are at least partly voluntary and reflect the agent's values on Shroeder's own account,) desires are not the only dispositions for which we can be held responsible. One illustration of this taken from Canadian law is Hundal vs. the Queen ([1987] 2 S.C.R. 1299), in which the defendant struck and killed another driver while driving an overloaded dump truck above the speed limit in dangerous conditions, and argued that he was innocent because he had not intended to cause any harm. The court found that his *de facto* intentions were irrelevant to the case, and chose to apply an 'objective' test of *mens rea,* based

on what a reasonable person would have done in the circumstances. This is a legal example, but Hundal's actions were clearly also morally blameworthy, regardless of the fact that he intended no harm. Thus, from a moral (as well as legal) perspective, it is irrelevant whether Hundal was driving recklessly due to habit, or whether his actions were 'the product of brute operant conditioning.' Specifically, the fact that his fatal course of action was due to operantly conditioned behaviour is morally irrelevant, *pace* Shroeder. What matters, on scrutiny, is not the cognitive origins of his behaviour, but the fact that, somehow or other, he should have known better. This shows that it is not only our desires for which we can be held accountable, but also for failing to have the desires that we should have had, given our capacities and circumstances.

This does not mean that Tourettic individuals are responsible for their tics. Rather, it implies that responsibility depends on whether an action could have been *avoided*, not whether it stemmed from, or reflected, a conscious desire. Tourettic individuals should be excused for ticcing because, or insofar as, their tics stem from a defect of the basal ganglia *over which they have no control.* Whether tics counts as 'motives' is irrelevant, except insofar as motives are under our control. Neurobiological studies of the basal ganglia indicate that defects to this mechanism result in a severe deficit of volitional control. On these grounds, even if we do not know the full extent of the cognitive damage, it is reasonable to assume that Tourtettic tics are not things for which agents can be held responsible. The fact that there may be some small degree of control is of little significance: just as it is physically possible for a person of extraordinary psychological fortitude to resist waterboarding torture, it may be possible for a person of extraordinary psychological fortitude to suppress Tourettic tics for a significant period of time; but this is not the standard to which we hold ordinary people. Ordinary people are held to a standard that is reasonably achievable for a person of average fortitude. On this view, Tourettic individuals are not responsible for their ticcing behaviour.

Still, Shroeder raises an interesting question for any view of responsibility. He writes, "of course there are Tourettic individuals who use their disorder as license to act badly, but in many cases tics are not taken to be expressions of ill will, even if they are reliably harmful" (p. 114). How does the avoidability view deal with such cases? If there are Tourettic individuals who feign tics, they are in principle morally responsible for this behaviour. However, there is no way of discerning (from a third person perspective) whether a Tourettic individual is exercising due restraint. This highlights a limitation of the objective view, and of all responsibility theories. Nonetheless, since we know that *all Tourettic individuals* suffer from a scientifically identifiable, clinically significant defect of the basal ganglia and the motor cortex, we should excuse all victims, unless we have good reason to suspect that they are deliberately malingering or failing to practice due diligence. (Making this determination would be a matter of using observable data.) Since the reasons view is no better equipped to discern when a person is malingering, it does not fare any better at assigning responsibility in such cases. However, the fact that it relies on such

obscure data as first-person subjective reports to assign responsibility in *all cases* of apparent wrongdoing is a mark against it. Externalism is methodologically distinct in that it privileges observable data and, by shunning reductive internalistic descriptions, acknowledges the multiple realizability of disabilities in terms of their cognitive etiologies.

**7. A Note on Methodology**

This section was prompted by correspondence with a very meticulous and insightful reviewer, who asked about my methodological leanings. The theory that I have proposed in this chapter is a pragmatic one. One of its merits, I believe, is that it provides clearer guidelines than internalist theories for when it is reasonable to hold a person responsible. This might not be seen as a desideratum if one is satisfied with abstract conceptual analysis, but most people want their philosophy to be applicable to real life. There are, I submit, several good reasons for preferring a pragmatic, action-guiding method. First, it stands to reason that responsibility ascriptions should be motivationally efficacious. To hold a person responsible for something that she could not have avoided, I have said, amounts to browbeating, a cruel and gratuitous practice. Indeed, there is growing consensus in the literature that the purpose of responsibility ascriptions is to direct behaviour. D. Copp (2008), for instance, contends that, "moral requirements are action-guiding. That is, their point is to guide agents' decisions among their alternatives… Given this, an adequate theory would imply that an agent is all-in morally required to do A only if she can do A" (p. 71). And Nelkin (2011) argues that moral responsibility conceptually entails 'action-directedness,' which explains why we can say that "rocks… do not have any [obligations]," but persons so (p. 115). I also hold to this requirement.

**8. Response to Objections**

In the preceding sections, I defended the use of SRP for diagnosing mental and neurological disorders. This theory is meant to be applicable to all cases of moral responsibility assessment, but it is particularly apt for defeating skepticism about moral responsibility in cases of mental illness. This is because it has three distinctive advantages over the alternatives, i.e., the *reasons view* and the *reflective self view*. Namely, (i) it avoids reliance on subjective data, which research in situationist psychology and neuroethics has shown to be elusive, unreliable, and relatively opaque to second-hand discernment (see, e.g., Doris 2002 2007, Harman 1999, Gazzaniga 2006), (ii) it accommodates our deep-seated intuitions about cases of negligence and peculiarly unfortunately formative circumstances, and (iii) it avoids depersonalizing the subject by exaggerating her incapacities, such that she is depicted as an object of treatment (p. 78). I see this view as a natural departure from earlier internalist accounts of moral responsibility, as

articulated by Frankfurt, Watson, and Taylor in the 1970s (i.e., the reflective self view), and Susan Wolf in the 1980s (i.e., the reasons view.) Just as Holmes' theory emerged to account for cases of culpable negligence, which traditional legal theories were incapable of explaining with subjective *mens rea*, moral philosophy must evolve, I believe, to account for these kinds of situational excuses. Holmes' theory was first used in an English tort law case in which the defendant, Menlove, had built a hayrick near his neighbor's property, in spite of warnings that it was likely to catch fire. When the structure ignited and destroyed his neighbor's cottages as predicted, Menlove was charged with culpable negligence. The courts deemed that regardless of his subjective intentions, Menlove should have known better.[10] So far in moral philosophy, there is no equivalent criterion which would allow us to judge a person's behavior irrespective of his subjective psychology. This is why moral philosophers (e.g., Scanlon 1986, Watson 1987, and arguably Wolf 1987) have struggled with the problem of responsibility-imputing negligence in the absence of ill will, and extenuating formative circumstances, which excuse transgressions even if the agent is evil or morally deranged.

This explains my main motive for developing an externalistic account of moral responsibility. Now, Perring and Bedrick, in personal correspondence, have raised some very insightful criticisms of this view, which compel me to clarify and enlarge upon this motive. There are three objections in particular that I see as warranting a fairly lengthy reply. The first is that we do not need, and perhaps cannot formulate in intuitively plausible terms, a coherent theory of moral responsibility. (This echoes Cartwright's skeptical worry.) The second is that an externalist account of moral responsibility is no more practicable than traditional internalist views. And the third is that SRP is not especially effective at preventing the depersonalization of psychiatric patients, and, moreover, we may sometimes have good reason to completely withdraw basic empathy and moral consideration. In what follows, I provide answers to each of these concerns.

Perhaps the most resonant objection is Perring's contention that we do not need, and perhaps cannot formulate, a coherent account of moral responsibility. Perring says,

> whatever philosophical theory we come up with is never going to achieve anything like certainty, and so we may come to the conclusion that in a large number of cases, either we can never know to what extent an individual is morally responsible, or to go one step further, there simply is no determinate answer to the question of whether a person is morally responsible in many of the difficult cases.

This concern is reiterated by Bedrick, who expresses dissatisfaction at my less-than-conclusive treatment of the Harris psychopathy case. My first reaction is to say that I never meant to imply

[10] *Vaughan v Menlove* (1837) 132 ER 490 (CP)

that SRP can adjudicate all cases of moral responsibility with absolute certainty. I do not believe that any moral theory is capable of eliminating uncertainty, or obviating the role of informed discretion in applying general principles to particular cases. In other words, moral reasoning necessarily involves both uncertainty and discretion. However, I do believe that my account is more practicable, in the sense of being more explanatorily robust as well as theoretically adequate than the alternatives, for two main reasons. First, SRP dispenses with subjective assessment, which situationist research shows to be more tenuous than objective, quantifiably data that can be interpersonally verified. This contention, of course, hinges on the viability of situationist psychology, which is by no means indisputable, but, by virtue of the sheer volume of corroborating evidentiary support, warrants serious consideration. In addition to examples already given, Gazzaniga (2005) highlights a number of cognitive distortions that undermine the credibility of first-person testimony, including memory fading, absentmindedness, blocking, misattribution, suggestibility, consistency bias, change bias, and egocentric bias (pp. 126-39). Perhaps the most well-known example is egocentric bias, which describes our tendency to overestimate our abilities and personal virtues. We can easily see how this bias could lead criminals to dismiss or rationalize their crimes, giving rise to unreliable (but honest) self-reporting. This lends weight to the notion that our introspective access, and *a fortioti* second-person testimony, tends to be remarkably unreliable. SRP urges us to distrust self-reports, and instead seek observable evidence of excusing conditions in either the agent's cognitive economy or local environment. Character theories do not explicitly advert to any particular methodological scruples, but they strongly suggest that introspection is methodologically natural and adequate. Returning to Frankfurt's influential account (1971), he says that an agent 'wills' an action if she has an "effective desire—one that moves (or will or would move) a person all the way to action" (p. 325). This is a very nuanced discrimination to make, and only one of several psychological criteria for evaluating moral responsibility. (A person must also possess higher-order volitions that correspond with the person's will.) The natural assumption is that 'willing,' 'wishing,' and 'having a volition,' which are three different psychological conditions, are simply introspectively available to the agent, and Frankfurt says nothing to counteract this presumption. Moreover, Frankfurt (1969) denies the 'principle of alternate possibilities,' which holds that a person is not responsible for her action if she could not have done otherwise, based on an ostensible counterexample, in which a person is psychologically manipulated by an evil neuroscientist to commit a murder, but the person also happens wholeheartedly to want to commit the murder. So the person's autonomous will and the evil neuroscientist's hopes coincide. But this is a very convenient example of wrongdoing. Frankfurt does not consider examples such as Ishamel Beah's, in which the person's whole motivational system was conditioned by the coercive influences of a corrupt person, in which case it is reasonable to say that person lacks responsibility, not only for his murderous actions, but for his very will. A similar example cited by Fischer (2002) is a 'Walden Two' scenario

(i.e., B.F. Skinner's fictional utopia, in which children are inculcated with communitarian belies and attitudes via rigorous operant conditioning.) According to Fischer, the inhabitants of this world are not responsible for their operantly conditioned character or consequent choices. Now, unlike Frankfurt, Fischer provides *two* criteria of responsibility, one psychological (internal), and the other historical (external); but, as Sneddon points out, he does not defend the second condition, and seems to want to reduce it to an internal capacity (More on this in chapter 6.) This fuels a presumption that psychological criteria are adequate for assessing responsibility, and that introspection is a sufficient methodological approach to ascertaining the relevant criteria. But this cannot be the case, since the agent's historical circumstances are the only true excusing factor in the scenario, given that nothing in the agent's psychological profile confers any exculpating force whatsoever. So this view is explanatorily inadequate, as well as methodologically misleading, since there is no reference to quantifiable methodological criteria, and suggestions to the contrary. My description of SRP, by contrast, explicitly requires such evidence. This methodological condition is necessitated by the situationist and neurological evidence.

The second reason for favoring SRP is that, unlike internalist theories, it is capable of accounting for externalistic types of inculpating and exculpating factors, such as culpable negligence, cultural isolation, and deprived childhood circumstances. If we consider American common law, negligence can be established irrespective of the defendant's subjective intentions; judges and prosecutors are permitted to consider cultural factors when charging someone with a crime, deciding appropriate sentencing, and plea bargaining with the defense attorney; and circumstances can inform judgments of insanity (HLRA 1986, p. 1295). According to Perring, "the law tends to set the bar high for any claim of diminished responsibility"; but if this is the case, it implies that these types of excuses should have even more force in moral reasoning than in legal reasoning. Yet moral philosophy has no generally-accepted equivalent to cultural and circumstantial excuses, which prevail in legal theory and practice. If Cane (2002) is right, moral philosophers can learn from legal theory, and use fine-grained legal considerations to refine moral responsibility.

Perring's second objection is that we may not need to justify our moral practices at all:

> When it comes to deciding whether someone they know is morally responsible in a wide range of cases, people often simply avoid the issue, live with uncertainty, or even simultaneously adopt two or more apparently incompatible views. It evidently is not necessary for us to resolve issues of moral responsibility in order to structure our social relationships and maintain social order… [Furthermore,] the suggestion… that philosophy could have a central role in the regulation of social relationships is in need of clarification.

Now, while I can understand Perring's skepticism about the practical utility of academic philosophy, I would contend that *philosophy* in its truest sense (and especially *moral* philosophy) is not merely, or even primarily, an academic discipline. Rather, to quote Lewis Vaughan in his 'Student's Guide to Writing Philosophy' (2006), it is "a field of inquiry concerned with the examination of beliefs of the most fundamental kind—beliefs that structure our lives, [and] shape our worldviews" (p. 3), a practice that entails "systematic, analytic, productive thinking," which is "useful in everyday situations" (p. 7). If we accept this definition, it is hard to see how one can live a good life without engaging in philosophical reasoning, or, conversely, how failing to engage in philosophical reasoning could not fail to undermine one's ability to live well. While it may be true that most people ascribe responsibility willy nilly, i.e., without any justification, or even any attempt at justification, these people should, ethically speaking, possess *some* reason for acting as they do. This is especially true when their moral practices affect other people's interests. For example, if a doctor is deciding whether or not to involuntarily hospitalize and forcibly medicate a patient, on the supposition that the person is morally impaired, the doctor has an obligation to provide justifying reasons to the patient, the patient's family, and the broader community; and given the doctor's considerable influence and authority, the reasons had better be pretty good. (This is why Canadian doctors, for example, are legally required to provide a notice justifying a judgment of incapacity.) While it is true that the physician may have developed a habit of relying on her medical intuitions, she cannot offer these intuitions as a *reason,* a *justification,* for revoking a person's civil liberties. She must offer interpersonally acceptable reasons. This is why we need a theory of responsibility that is empirically robust and fairly intuitively grounded: because as moral agents, we have a responsibility to justify our other-oriented actions to those whom they affect, and to the broader community of humans who have a stake in the social good. Admittedly, a medical doctor is a particularly high-profile case with a particularly high burden of responsibility, but the same basic considerations apply to ordinary people. For instance, if a mother is going to withhold affection and moral consideration from her son, she owes him and moral society an explanation.

That said, I believe that people tend to have fairly cohesive moral intuitions, and that, generally speaking, we are interested in acting on justifiable, internally coherent beliefs and attitudes. This is why we tend to balk at the accusation that we have been hypocritical or irrational, rather than simply shrugging it off, or failing to appreciate its force. This is one of my reasons for maintaining that an adequate moral theory should be able to accommodate our pretheoretical intuitions and considered judgments, ideally integrating them in a 'reflective equilibrium.' I believe that SRP has an advantage in this regard, because it is, on my view, a deliberative heuristic, as opposed to a reductivist theory. That is, like Rawls' *original position*, it provides a reflective method for justifying our moral practices based on multiple considerations, rather than positing grounding principles, or necessary conditions, of moral agency. This aspect

respects the interpersonal nature of moral responsibility, which seeks to ascribe responsibility on the basis of public justifications.

This brings me to the third objection, i.e., that SRP is not especially effective at preventing depersonalization, and that depersonalization may even be warranted. To begin, I should explain what I take 'personhood' to mean. Of all the theories I have encountered, I favour Mary Anne Warren's (1973) description of *personhood* as involving six basic capacities: sentience, emotionality, reason, the capacity to communicate, self-awareness, and moral agency. Of these conditions, "none may be logically necessary, but the more criteria that are satisfied, the more confident we are that the concept is applicable" (Warren, p. 131). This definition excludes fetuses and unrecoverable coma patients from personhood, but encompasses most adults, including those with fairly severe cognitive deficits. Since the criteria are not individually necessary, a person might lack a particular capacity—such as moral agency—and yet still qualify as a person. Thus, on this model, there is no warrant for excluding psychopaths from personhood, since they may have, at least, sentience, the capacity to communicate, and self-awareness to a very high degree, and possibly also some degree of emotionality and reasoning. The only reason that I can see for excluding them is if we equate personhood identically with moral agency, but to reduce personhood to this single capacity needs an argument, and one that defeats Warren's sound reasoning. Such an account may be possible, but most philosophers seem to take this reduction for granted, rather than defending it. (Kant, for example, is famous for reducing personhood to rationality.) In addition, it is unclear what a reductive approach would accomplish, since excluding individuals from interpersonal relationships deprives them of any motive to act morally. Thus, even if there is reason to think that someone is severely deficient, it is simply more prudent, and more humane, to assume that the agent might be able to achieve a degree of interpersonal engagement if afforded sufficient encouragement, opportunities, and resources, and to treat her as a person on tentative prospective grounds. On this point, it is worth noting that Western psychiatry has perpetuated horrible abuses against psychiatric patients, and arguably still violates people's basic rights, and this is largely attributable to the industry's depersonalizing attitude toward patients (see, e.g., Bentall 2004, Graham 2010, and Greenberg 2013). This gives us especially strong reason presumptively to treat psychiatric patients as persons rather than objects. Now, while nothing that I have said logically rules out objectifying psychiatric patients, my arguments at least place a very strong burden of proof on anyone who wishes to take this stance.

### 8. Concluding Remarks

This chapter has argued that SRP is preferable to the reflective-self view and the reasons view for evaluating moral responsibility in cases of mental health disabilities, as it answers adequately to Cartwright's main objections. It is also preferable in non-mental-health cases, because it avoids the epistemological problem, and captures the broadest range of intuitive excusing conditions. If we examine the history of the philosophy of moral responsibility, the reflective-self view emerged in the 1970s in the writings of Frankfurt, Watson and Taylor, and the reasons view emerged in response to this position in the 1980s, largely thanks to Wolf's influential counterargument (1987). Wolf's view added to the reflective-self view the condition that a responsible agent must have minimal normative competence or sanity. I have now argued that in order to make sense of Wolf's own central example, we must add a further, external condition: namely, that a responsible agent must have access to relevant environmental resources throughout her personal development. This criterion is supported by scholarship in feminist epistemology (Fricker) and the philosophy of law (Holmes,) which demonstrates that internalist theories neglect compelling extenuating factors. In light of these shortcomings, we would do well to consider SRP as a strong contender.

## Chapter 6: Externalism as Method: Justifying Strawson and the Excuse of Peculiarly Unfortunate Formative Circumstances

### 1. Preliminaries

In this chapter and the next, I continue to defend methodological externalism, and describe it as a viable interpretation of Strawson's famous account of moral responsibility *qua* expression of the reactive attitudes. This situates externalism in the literature as a form of Strawsonianism, in opposition to alternative theories such as Frankfurt's, Watson's, and Scanlon's. My view is informed by Andrew Sneddon's interpretation (2005), which contends that philosophers have been mistaken in reading Strawson through an internalist lens, and in construing Strawsonian responsibility in terms of internal, psychological capacities. Defining responsibility in this manner supports a mechanistic, deterministic view of agency, which Strawson explicitly rejects as "external to the participant perspective within which the reactive attitudes are deployed" (Sneddon, p. 246). Sneddon recommends that we conceive of moral responsibility instead as a social competence, evidenced in our ability to interact appropriately with others. In addition to being external in the sense of *interpersonal* (and thus 'metaphysically wide,') this view is also external in defining competence in terms of our *overt behaviour*, rather than our internal, psychological profile. I use this interpretation as a means of fleshing out my interpersonal theory

in broadly Strawsonian terms, but my argument is not purely derivative upon Sneddon's. I enlarge on Sneddon's account by educing new implications therefrom. In particular, I demonstrate that the externalist approach solves a longstanding problem in moral philosophy, which is the question of how to support Strawson's excuse of peculiarly unfortunate formative circumstances (henceforth, PUFC,) which also appears, in different words, in Wolf's and Fricker's works, amongst other places. Since this excusing condition is situational and external, i.e., outside of the agent, it is unclear how an internalistic theory of responsibility could grant it even partial legitimacy. I show that an externalistic interpretation of Strawson can accommodate this excuse, and thus accommodate Strawson's roster of excusing conditions. At the end of the chapter, I very briefly offer several other reasons for preferring Sneddon's interpretation, which have already, to some extent, been demonstrated in this dissertation. First, I say that Sneddon's view accommodates findings in social psychology and cognitive science which call into question our introspective access, and which show psychology-based theories to be impracticable. Secondly, I draw an analogy between moral deficits and learning deficits to show that psychological theories of responsibility are inadequate insofar as they postulate a single, identifiable cognitive deficit as *the* source of moral incompetence/lack of responsibility. Trying to reduce complex behavioural deficits to single cognitive deficits overlooks the fact that human being are remarkably adept at compensating for internal defects by recruiting and utilizing alternative cognitive resources, as exemplified in the case of dyslexia. Thus, a type-type reductive theory of moral deficits to cognitive deficits is misguided. More than one cognitive resource can be pressed into moral service.

In chapter 7, I continue this methodological argument using research from Faraci and Shoemaker (2013), which, on my interpretation, shows that externalistic excuses such as PUFC figure in the intuitions of naïve moral reasoners. This lends credence to Wolf's premise that this type of excuse is embedded at the level of 'our shared pretheoretical intuitions.' Although Faraci and Shoemaker offer a different interpretation of the survey results, I argue that this construal is implausible given that PUFCs were the only significant differential variable across three scenarios presented to the subjects. The researchers attempt to explain away this result by imputing sophisticated rationales to their subjects, but I do not believe that these motives can legitimately be attributed to naïve reasoners, who are supposed to be philosophically uninformed. I urge that we take the survey results at face value, which means abandoning the researchers' preferred account, i.e., the deep-self view, and adopting an externalistic framework. If this interpretation is correct, it corroborates the externalistic account of responsibility defended in previous chapters.

**2. Introduction**

P.F. Strawson's theory of moral responsibility remains eminently influential, and has been the

starting point for many post-Strawsonian discussions of responsibility. Strawson defines moral responsibility in terms of our tendency to express the reactive attitudes of gratitude, resentment, approbation, disapprobation, and so on, toward others in light of their intentional actions. This theory is supposed to be justified by "human nature and our membership of human communities" (Strawson, p. 1985). Since the publication of Strawson's seminal work (1963), most moral philosophers have found this theory intuitively appealing, but many have disputed its interpersonal justification, and demanded a stronger explanatory rationale. Watson (1987) and Scanlon (1986), for instance, have argued that Strawson does not adequately explain when it is (normatively) appropriate to modify or suspend the reactive attitudes. To resolve this problem, they have provided an expressivist and a contractualist interpretation, respectively. However, one cannot help but notice that these interpretations exclude one of the criteria that Strawson cites under the rubric of exempting conditions: namely, the excuse of being "peculiarly unfortunate in formative circumstances" (p. 79). For my part, I do not think that we can so easily jettison this excuse, especially considering that other philosophers have defended some version of it (e.g., J.M. Fischer, S. Wolf), albeit somewhat incoherently, and it seems to draw intuitive assent. Moreover, we should hesitate to dismantle Strawson's repertoire so quickly.

In this chapter, I argue that Watson and Scanlon are right that we need a stronger normative account of when it is appropriate to deploy the reactive attitudes, but their theories are inadequate. Furthermore, Wolf's sane deep-self view, which *prima facie* seems to preserve the condition of PUFC, is too vague, and rests on a questionable premise. However, Sneddon's little-known externalist interpretation of Strawson provides an avenue for progress. Sneddon says that we should understand moral responsibility as a social competence similar to playing tennis, and that we should assess this type of competence based on the agent's overt beahaviour as manifested in her interpersonal interactions, rather than grounding it psychological mechanisms. Since social competency is demonstrated interpersonally as opposed to individualistically and psychologically, Sneddon's theory accords with Strawson's injunction that we must justify the reactive attitudes from *within the participant perspective*. In my final section, I argue that the externalist interpretation supports PUFC, through the posited analogy between moral competence and other types of normative competences (such as tennis, boxing, linguistic ability), which entails that we should excuse moral outsiders or ingénues. That is, just as we would excuse a novice tennis-player for not performing at a professional level, we rationally ought to excuse ethical novices (i.e., those who have not undergone a normal process of enculturation) for failing to achieve the normal level of moral competence. This means that if someone has been denied access to moral perspectives due to PUFC, it is reasonable to partially or fully excuse the person from responsibility. However, the degree of extenuation must correspond to the degree of epistemic constraint, as evinced in the person's social location. In modern western democracies, this degree will be fairly low due to the ubiquity of information and the multiplicity of perspectives.

### 3. Strawson, the Reactive Attitudes, and the Excuse of Peculiarly Unfortunate Formative Circumstances (PUFC)

On Strawson's view, regarding people as morally responsible consists in our tendency to express the reactive attitudes, which are rooted in human nature and our interpersonal relationships. Thus, "questions of justification are internal to [their] structure" (p. 91). However, we might be expected to modify or suspend these attitudes under two conditions: first, if a person "didn't mean to cause harm," "wasn't himself," or "was acting under post-hypnotic suggestion," (Strawson, p. 79); and second, if an agent appears to be immature, "psychologically abnormal," "morally undeveloped," or "peculiarly unfortunate in his formative circumstances" (Strawson, p. 79). (I have quoted these conditions verbatim, as elaborating on them at this point would bias the discussion in favour of a particular theoretical interpretation.) We might call the first class of constraints *excusing conditions* and the second, following Watson, "*exempting conditions*," since the latter conditions are supposed to completely exempt the agent from participation in human relationships. Toward the second group of agents, we are supposed to adopt the "objective attitude," whereby we view the individual "as an object of social policy…, a subject for… treatment" (Strawson, p. 79). Since the reactive attitudes are essentially human 'givens,' "the general framework of attitudes itself…, as a whole…, neither calls for, nor permits, an external 'rational' justification" (Strawson, p. 91). Those who seek a deeper explanatory rationale are said to be "over-intellectualizing the facts" of moral life (Strawson, p. 91).

Since the publication of Strawson's paper, most moral philosophers have agreed that his extenuating conditions are intuitively plausible, but many have disputed their justificatory basis, ignoring the charge of over-intellectualization (viz., Watson 1987, Wolf 1987, Shoeman 1987, and Scanlon 1986, 1988). Indeed, from an impartial philosophical perspective, an appeal to the interpersonal sentiments as a basis for moral responsibility can seem, at best, insufficient grounding for a normative theory, and, at worst, a form of emotional appeal or *ad populum* fallacy. Watson and Scanlon present what I take to be two of the strongest criticisms of Strawson. Watson contends that, "Strawson's theory does not provide an account of how [our proneness to exempting conditions] works or what kinds of explanations exempt" (p. 263). This results in the theory being not only "incomplete" but incoherent: "what might be necessary to complete it will undermine the theory" (p. 263). The theory is incomplete because its simplistic description of exempting conditions fails to provide a non-question-begging explanation of why it is (normatively) appropriate to deploy the objective attitude toward those deemed psychologically 'deranged.' It is incoherent insofar as a more robust explanation would invoke rationales that go against Strawson's over-intellectualization objection. The need for this type of explanation, however, is exigent, given that the various extenuating conditions do not clearly fit together in a

coherent whole (at least from an internalist perspective.) The excuse of PUFC in particular seems to resist integration, as it rests on historical factors which, Watson says, "can be, at most, evidence that some other plea is satisfied" (p. 274). That is, Strawson's account does not spell out how or why historical factors should make a (non-derivative) difference to our reactive attitudes. Moreover, an adequate defense is not obviously available in moral philosophy in general.

Watson does not think that these flaws are fatal. He believes that we can salvage the core of Strawson's theory by grounding the reactive attitudes in the interpersonal exchange of moral demands. Specifically, we should regard the reactive attitudes on an expressivist model, as speech acts "expressing a moral demand, a basic demand for reasonable regard" (p. 264), and we should see extenuating conditions as "constrains on intelligible moral demand, or… moral address" (p. 236). This interpretation explains why we tend to exonerate, for example, young children and psychopaths, i.e., because they are not fully capable of grasping moral concepts and engaging in moral communication (Watson, p. 265). The one notable exception to Watson's explanatory framework is PUFC, which cannot be directly justified on expressivist grounds. At most, this condition can be seen as *evidence* of an expressivist defect. Watson notes this problem, and considers three possible alternative, *ad hoc* justifications: (i) that victims of PUFC are especially deserving of sympathy, (ii) that we would have become as vile as the perpetrator under the same circumstances, and (iii) that an unfortunate childhood necessitates certain behaviors in such a way as to excuse the individual in question. Watson deems all of the explanations unpersuasive, and so he rejects PUFC.

Similarly, Scanlon finds Strawson's theory generally "plausible and appealing" (p. 165), but he concurs that it fails to explain how and why exempting conditions should influence the reactive attitudes. He notes that Strawson's theory "can be understood on two levels" (p. 162): (i) as a descriptive account of moral psychology and our proclivity to experience the reactive attitudes, or (ii) as a normative account of when and why it is *appropriate* to express these attitudes. Unfortunately, neither of these interpretations is adequate, since the first one is non-normative and the second one is question-begging. Furthermore, "it is not clear that moral judgments need always involve the *expression* of any particular reactive attitude" (p. 165):

> For example, I may believe that an action of a friend, to whom many horrible things have recently happened, is morally blameworthy. But need this belief, or its expression, involve a feeling or expression of moral indignation or disapproval on my part? Might I not agree that what he did was wrong but be incapable of feeling disapproval toward him? (p. 166)

Following these objections, Scanlon introduces his own theoretical substructure for moral

responsibility. He purports to explain Strawson's theory by grounding the reactive attitudes in "moral reasoning and moral motivation" (p. 167), as per his contractualist view. Specifically, he says that it is appropriate to suspend the reactive attitudes toward persons who are incapable of negotiating and engaging in moral reasoning. This explanation supports Strawson's claim that "moral judgments presuppose a form of interpersonal relationship. On the contractualist view, moral judgments apply to people considered as possible participants in a system of codeliberation" (Scanlon, p. 167). It also divorces the reactive attitudes from their overt expression in words and behaviour, by defining moral fault as an (objective) departure from contractualist principles, i.e., principles that no suitably-motivated person could reasonably reject. Finally, it avoids question-begging assumptions because it is "essentially cognitivist," i.e., it "can explain why moral judgments would normally be accompanied by certain attitudes, but these attitudes are not the basis of its account of moral judgment" (Scanlon, p. 167). Thus, it patently comports with Stawson's theory in many ways.

At this juncture, it is important to note that Watson and Scanlon purport to be *explaining* Strawson's theory, rather than providing a substitute or substantial modification of it. This can be seen in Watson's claim that he intends to "remedy" the "incompleteness" of the theory "in Strawsonian ways" (p. 275) and "in Strawsonian terms" (p. 264). Scanlon similarly sets out to "sketch briefly how a Quality of Will theory [such as Strawson's] might be based on a contractualist account of moral judgment" (p. 166), and to show how "contractualism gives specific content to the idea, suggested by Strawson, that moral judgments presuppose a form of interpersonal relationship" (p. 167). Scanlon and Watson diverge from Strawson, however, in rejecting the excuse of PUFC, which expressivism and contractualism cannot explain. While this divergence is not perfectly transparent in Scanlon's *Tanner Lectures*, he goes so far as to say that while we can excuse agents who lack critically reflective, rational self-governance, "if what is 'lost' is more specifically moral—if, for example, a person lacks any concern for the welfare of others—then the result begins to look more like a species of moral fault" (p. 27). This implies that individuals cannot be excused for moral faults or incapacities due to *formative circumstances*, or on any other causal grounds. Faraci and Shoemaker confirm this interpretation of Scanlon by classifying him along with Angela Smith as a "new" protagonist of the "deep self view," "a label coined by Susan Wolf to describe a view originally advocated (in different forms) by Frankfurt, Taylor, and Watson" (p. 320). This view differs from Wolf's in that it does not excuse morally insane individuals from blame to any degree.

Hence, we see that while Watson and Scanlon offer different theoretical interpretations of Strawson, neither can accommodate the excuse of PUFC. This excuse is interesting and important in its own right, as it has attracted a great deal of philosophical attention and scrutiny (viz., Watson, Wolf, Faraci and Shoemaker, Shoeman.) Wolf believes that this condition—which she describes as "deprived childhood circumstances" and "related cases" (p. 382)—is rooted in

"our pretheoretical intuitions" (p. 382), as well as embedded in our social, political, and legal institutions, such as the M'Naughten Rule, which excuses persons from legal responsibility if they do not understand the difference between right and wrong (Wolf, p. 82). If Wolf is right, then we have *prima facie* reason to try to preserve PUFC as a morally relevant consideration. However, I agree with Watson and Scanlon that a non-question-begging normative account of when and how we should deploy the reactive attitudes is needed if we are to overcome our reliance on emotional appeals and *ad populum* reasoning. As Wolf's view appears to depend heavily on a single intuitive example—what Denett calls an 'intuition pump'—it seems vulnerable to exactly this type of objection.

In the next section, I examine Wolf's account of moral responsibility, the sane deep-self view. I begin by outlining Wolf's sanity condition and her famous JoJo example, in more detail than previously attempted. I then argue that it is not clear if PUFC is supposed to be inherently excusing or merely evidential on Wolf's account, and if it is the former, Wolf provides insufficient argumentation for this interpretation. In section 4, I describe an alternative, *externalist* reading of the reactive attitudes, which provides a more robust explanation of Stawsonian responsibility, and allows us to retain Strawson's full repertoire of excuses.

**4. Moral Insanity and Deprived Childhood Victims**

Wolf is well known in moral philosophy for her defense of the excuse of moral insanity. In *Sanity and the Metaphysics of Responsibility*, she also mentions the condition of "deprived childhood circumstances" (p. 382), but it is not immediately evident whether this latter condition is supposed to be independently extenuating, or merely a *prima facie* indicator of an intrinsically extenuating deficit within the agent's brain or mind. If it is merely an indicator, the theory is vulnerable to Watson's objection that "in themselves, [facts about backgrounds] do not seem to matter" (p. 274). This condition is, of course, centrally important for our purposes, as it reiterates, in slightly different language, Strawson's condition of PUFC. In response to Wolf (2010), Faraci and Shoemaker contend that neither moral insanity nor deprived childhood circumstances are excusing from a lay perspective, and so, "Wolf's assumptions about our pretheoretic intuitions… are wrong" (p. 330). They draw this conclusion based on a survey of their students at Bowling Green State University. Although this is merely anecdotal, it is interesting that when I offered the same survey to my own undergraduate students, I got quite different results: the majority were inclined to excuse deprived childhood victims.

There is insufficient evidence to settle the empirical debate at present, so let us evaluate Wolf's arguments on their own merits, and consider whether, in philosophical terms, they justify an excuse of PUFC. Wolf holds that in order to be morally responsible, an agent must satisfy three conditions: (i) her actions must be within the control of her will, (ii) her will must be within the

control of her deep self, and (iii) her deep self must be 'sane'—that is, it must "have the [general] ability cognitively and normatively to understand and appreciate the world for what it is" (p. 387). This view relies on the premise that our pretheoretical intuitions support this type of excuse, which is supported by the following thought experiment:

> JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, it is not surprising that little JoJo takes his father as a role model and develops values very much like Dad's. As an adult, he does many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things, he acts according to his own desires. Moreover, these are desires he wholly wants to have. When he steps back and asks, "Do I really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life expresses a crazy sort of power that forms part of his deepest ideal. (p. 379)

Wolf concludes that, "in light of JoJo's inheritance and upbringing… it is dubious at best that he should be regarded as responsible for what he does" (p. 380). Furthermore, "it is unclear whether anyone with a childhood such as his could have developed into anything but the twisted and perverse sort of person he has become" (p. 380).

Now, while I personally find this example intuitively compelling, this is beside the point. If we aim to take seriously Watson's and Scanlon's criticism, we need a stronger philosophical basis for our judgments than appeals to intuitions and intersubjective agreement. We cannot, in other words, let the debate boil down to a battle of intuitions. A second point of contention is that it is not clear whether the sane deep-self view is even supposed to accommodate PUFC-type excuses. It is difficult to see how it should not, since deprived childhood victimhood, and other environmental references, figure prominently in Wolf's article, and they seem to be doing a lot of the intuitive work in the JoJo example—particularly the stipulation that JoJo is from a 'small, undeveloped country,' which suggests that he could not have known better. However, the *sane-deep self view*, in its name and its broader theoretical articulation, strongly suggests that moral insanity is the relevant factor, while external circumstances are merely evidential.

For these reasons, I want to consider an alternative account of moral responsibility. In the next section, I introduce an underappreciated methodological approach advanced by Sneddon (2005). This approach is an *externalist, pragmatic* one, which evaluates moral responsibility on the basis of overt behavioural criteria (namely, behavioural indicators of social competency), as opposed to individualistic, psychological ones (such as cognitive structures.) This reading

supplies explanatory resources for an alternative interpretation of Strawson, which both fits better with his view of the participant perspective, and legitimates PUFC.

**5. Moral Responsibility as a Social Competence: Saving PUFC**

Sneddon introduces a completely novel interpretation of the reactive attitudes which, as far as I can tell, has not been defended anywhere else in moral philosophy, except perhaps in my own work (2010). I developed a theory similar to Sneddon's through independent research, and only later discovered his similar argumentation (2005). Sneddon recognizes that externalism is original and widely "overlooked" (p. 239):

> So far as I can tell, *almost all* major contemporary theorists pursue the individualistic approach (Watson, 2001; Fischer, 1999). That is, they all theorize about moral responsibility in such a way that changes solely to an agent's environment and not to the intrinsic properties of the agent cannot affect the agent's status as morally responsible. (p. 242, emphasis in original)

Sneddon cites John Martin Fischer and Mark Ravizza, R. Jay Wallace, and Philip Pettit as examples of internalism, but the same interpretation can be applied to Watson, Scanlon, and Wolf. By contrast, "Strawson's own position, developed and extended to the issue of being morally responsible, is externalistic. Given the influence of Strawson's position, it is very interesting and peculiar that contemporary work that is explicitly influenced by Strawson diverges from his position in this way" (Sneddon, p. 242).

I concur with Sneddon, and I hope that this citation will bring more attention to his interesting research. Since Sneddon deals with Strawson specifically, I shall elucidate his interpretation of the reactive attitudes, and then highlight an important implication which he does not explicitly draw, regarding PUFC. Namely, I will show that externalism provides a normative foundation for this excuse. I will also recapitulate and reinforce Sneddon's defense of externliasm as an approach that respects the participant perspective.

To begin, Sneddon makes two main claims, one formal and the other methodological. First, he says that Strawson's notion of moral responsibility must be understood as a social competence. This is because for Strawson, "to be morally responsible is to be an apt candidate for the reactive attitudes. Put another way, to be morally responsible is to fit into the social practices governing the deployment of the reactive attitudes. In short, it is to acquire a social competence" (p. 241). In effect, Sneddon believes that internalism is incompatible with Strawson's insistence on employing interpersonal explanatory criteria, whereas internalistic explanations rely on psychological data, such as cognitive and psychological capacities, to justify

the reactive attitudes. But these are causal explanations, which *ipso facto* are external to the participant perspective. Within Strawson's framework, it makes more sense to construe moral responsibility as a social competence manifested in interpersonal relationships, as opposed to a psychological capacity internal to the agent. This has methodological implications. While psychological capacities may underlie moral behvaiour, they are not necessary and sufficient for responsibilty, nor are they introspectively available. Thus, we should utilize observable evidence, which is mandated by externalism.

Sneddon's second claim is strictly methodological. He argues that there are two ways of construing social competence: on an internalistic model and on an externalistic model. An "internalistically construed social competence," or *competenceI*, "can be explained solely in terms of the intrinsic properties of the individual agent," whereas an "externalistically construed social competence," or *competenceE*, "can be explained in terms of the way the agent fits into his/her context. Instead of solely using intrinsic properties of the agent, relational ones would also be used in the explanation of the competence in question" (p. 241). To illustrate the distinction between these two modes, Sneddon gives a set of parochial examples: tying one's shoelaces versus playing tennis. Shoe lace-tying, he says, is a *competenceI* because what counts as shoe lace-tying in one context is the same in every context in which the practice exists. If there is an idiosyncratic form of shoe lace-knotting which exists in some distant culture, it does not count on our conception. Playing tennis, by contrast, is a *competenceE* because it can be realized in many different ways. Taking the example of Andre Agassi, Andy Roddick, and Martina Navratilova, we find three different manifestations of tennis-playing competence: Agassi excels in his serve-returning ability, Roddick is an average returner but excels at serving, and Navratilova excels by employing a multifarious strategy consisting of serving, running up to the net, and controlling gameplay by delivering volleys rather than groundstrokes (Sneddon, p. 244). All three tennis players are commensurably competent, but they manifest their competence through different strategies, different cognitive processes, and different courses of action. This shows that *cometenceE* "does not require any particular abilities in order to be realized" (Sneddon, p. 245). By analogical induction, it follows that moral responsibility does not require any particular abilities in order to be realized: it can be instantiated through different strategies and different causal processes within the agent's psychological economy. All that is required is the ability to act "sensitively to the deployment of the reactive attitudes, and as if using these attitudes oneself": in fact, "whether one actually experiences the feelings seems to be beside the point, since whether one does or not is inaccessible to others" (Sneddon, p. 247). In this way, moral responsibility is a multiply realizable social capacity.

This view runs counter to internalistic theories that tie moral responsibility to particular, necessary psychological structures or capacities, such as second-order desires (Frankfurt), a valuational system (Watson), a suitably reasons-responsive mechanism (Fischer), or sanity,

defined psychologically as the *ability* to understand and appreciate the world (Wolf). Regarding Wolf's view, it may seem at first blush to be methodologically externalistic, since it includes a putatively externalistic criterion, i.e., deprived childhood circumstances. However, there is reason to reject this interpretation. Sneddon, when assessing Fischer's account, notes that Fischer explicitly specifies two criteria of responsibility: (i) a suitably reasons-responsive mechanism, and (ii) taking ownership over this mechanism. The second condition is supposed to imply that the agent has not been subjected to objectionable external manipulation, such as hypnosis, nefarious brain surgery, or Skinnerian conditioning. Having two types of conditions may seem unproblematic, but Sneddon identifies two worries. First, although the reasons-responsive condition is putatively internalistic, Fischer nowhere defends this methodological assumption. Thus his account is incomplete (Sneddon, p. 253). And secondly, while the 'ownership' condition is putatively externalistic, it "is more closely related than one might think to the search for individualistic psychological criteria for moral responsibility. Having identified a set of conditions that can undermine responsibility, Fischer seeks something specific about individuals to pre-empt the problem" (p. 254): namely, the absence of a suitably reasons-responsive mechanism that the agent has made her own. This means that the second criterion is reducible to an internalistic condition, rendering the externalist reading otiose. Thus, Fischer's view is susceptible to Watson's objection that historical conditions are at most evidentiary, if not completely irrelevant.

Now the same type of criticism can be leveled against Wolf. First, if the criterion of moral sanity is irreducibly internalistic, Wolf nowhere explicitly defends this methodological assumption. And secondly, the criterion of deprived childhood victimhood appears to be reducible to an internalistic criterion, making it susceptible to Watson's objection. Sneddon's account of Stawsonian responsibility, however, includes an explicit defense of externalism, rendering it immune to question-begging criticisms. To summarize, he argues that Strawsonian responsibility is best understood as a social *cometenceE*, analogous with tennis, as opposed to a *competenceI,* such as lace-tying. By contrast, internalism relies on internal causal criteria, which "have their natural home in the domain of determinist discourse, which is external to the participant perspective within which the reactive attitudes are deployed" (Sneddon, p. 245). However, externalism utilizes interpersonal evaluative criteria, which fits with Strawson's implied methodology.

Now that I have elucidated Sneddon's account and how it relates to Wolf, I am in a position to explain its connection with PUFC. Sneddon nowhere mentions PUFC, and it is not immediately obvious how externalism might justify this type of excuse. We cannot simply assume that because PUFC is situational and historical, it is *ipso facto* justified on an externalist model. We need a more precise explanation of this relationship. This can be accomplished, I believe, by examining the analogical relationship between moral responsibility and other externalistically construed social competences (viz., tennis.) If we revisit Sneddon's tennis example, we find

evidence that the demands of competency vary according to a player's experience. In arguing that a *competenceE* does not require any particular necessary abilities, Sneddon compares novice tennis players against experts, saying, "for beginner competence, what is required is to be able to get the ball over the net" (p. 245), whereas for expert competence, what is required is to be able to hold one's own against other experts. The point of this argument is to show that competence is not measured by or reducible to a person's internalistically construed abilities (Sneddon, p. 245). But we can also draw the conclusion that standards of competency vary from one person to another depending on the individual's level of experience, education, and practice. By parity of reasoning, this means that moral responsibility must vary according to a person's level of experience, education, and practice. We may reasonably infer, for example, that if a person has recently emigrated from a foreign culture with significantly different social customs, we are not justified in holding the person to the same standards as national-born citizens or landed immigrants who have inhabited the culture for more than 1095 days (or whatever timeline is required in the country at issue.) That is, just as it is reasonable to excuse a novice tennis player for performing below an expert's level of competence, it is reasonable to excuse a recent immigrant for failing to conform to local customs before a reasonable period of acclimation. We should expect this competence to develop in people gradually, and should thus hold cultural outsiders to increasingly strict standards until they reach the chronological status of landed immigrants.

There is support for this inference in Sneddon's account of moral psychology. Sneddon says that moral standards are more or less culturally and historically relative, within limits set by human nature. Specifically, there may be "universal psychological attributes of moral responsibility," but the expression of the reactive attitudes will be culturally, geographically, and historically sensitive (Sneddon, p. 262). This means, I think, that interpreting and responding to the reactive attitudes in a given context will require a degree of acclimatization and acculturation. It follows that we should not, for instance, hold a New Guinnean transplant to the same standard that we would normally apply to the average metropolitan westerner, since interpreting social customs requires a degree of familiarity and experiential exposure. Likewise, we cannot reasonably expect a 20-year old undergraduate student to show the same well-honed moral sensitivity that we expect of an ethics professor. This differential supposition seems eminently commonsensical, and coheres with our attitude in, say, sports, where we hold athletes to differential standards depending on their experience, height, weight, and so on. On close inspection, a deeper philosophical justification for this attitude can be found the principle of ought-implies-can, on the practical understanding that we employ in daily life. For instance, we do not expect children to solve complex algebraic problems, physically disabled people to compete against able-bodied people in the Olympics, or recent immigrants to speak fluent English. (Or, if we do, we are arguably not only unreasonable, but also bigoted.) It stands to

reason that we should not hold normal citizens and social outliers to invariant moral standards.

At this juncture, one might question the relationship between playing tennis competently and knowing the rules of the game. On my understanding, this is analogous to the relationship between knowing-that and knowing-how, to quote Gilbert Ryle (1949). That is, tennis competence is not reducible to propositional knowledge, and does not even require such knowledge as a minimal condition. Many, perhaps most, tennis players have never read a book on tennis, and, while discussing the rules of the game may be helpful, we can envision someone achieving adequate tennis competent without engaging in conversations about formal strategy. In this connection, it may also be instructive to note that Plato's conception of practical wisdom is, according to many virtue scholars (e.g., John Gould 1955), a form of Rylean know-how, and Aristotle draws a similar distinction between theoretical knowledge, which is propositional, and practical knowledge, which is essentially "concerned with action" (NE ch. 7), and evolves through habituation. This latter kind of knowledge is what I imagine to be involved in tennis competence, as well as, by analogy, the capacity for moral responsibility. It follows that in order to achieve this type of competence, one does not need to understand moral principles or rules *per se*, but one must develop dispositions to act in accordance with interpersonal practices, by actively participating in moral life. Notably, this comports with Haidt's understanding of moral reasoning as inherently inscrutable, which is why people typically cannot justify their moral intuitions. But this does not mean that their intuitions are wrong. Insofar as socialization is effective at conferring moral dispositions, critical reflection is neither sufficient nor necessary.

These considerations have favourable implications for PUFC. In Wolf's example, JoJo is supposed to be from a small, undeveloped country, which implies that he was never exposed to extraneous influences. We are also told that JoJo was given a "special education" from his doting father, and it is "not surprising" that he turned out to be a sadistic dictator (Wolf, p. 379). This again emphasizes his epistemic isolation. In light of these specifications, JoJo does not appear to be very different from the foreigner from a remote ethnic enclave who has not been exposed to extraneous moral influences, except for one thing: JoJo seems to be *even more* isolated than most conceivable immigrants in this day and age. In the context of the 'global village,' it is relatively easy for most people, even in developing countries, to gain access to alternative perspectives through technology and other forms of cultural dispersion. This explains why we are very reluctant to take an unmitigated excusing attitude toward *anyone*, including recent immigrants. We tend to assume that everyone has had some access to moral enlightenment, and thus if someone is morally ignorant, it must be at least partly due to culpable neglect. However, while this assumption is generally accurate, it is not unexceptionable: there are rare individuals who are *especially* or *inordinately* or *peculiarly* isolated from an early age. Toward these people, we tend to mollify our reactive attitudes significantly. This is reflected in the legal system, where judges may recognize "culture as a factor to be considered before imposing sentence" (HLRA, p.

1294), as well as the writings of Michael Slote (1982) and Gideon Rosen (2003). It is also found in Beah's autobiography. From an internalist perspective, it is difficult to justify this intuition, and I do not believe that Slote or Rosen provide non-question-begging arguments. However, this excuse becomes obvious when we construe moral responsibility on an externalistic model, as a social competence analogous with other competences. Then we are permitted to evaluate people according to their degree of experience, education, and acculturation, as we do tennis players, boxers, math students, etc. This does not mean that we should *completely suspend* the reactive attitudes in cases of PUFC. Rather, the degree of suspension should reflect degree of cultural isolation. JoJo's case may warrant complete suspension, while a moral transgression from a 21st century cultural transplant may warrant partial suspension, depending on the person's degree of seclusion. This is formally consistent with Strawson's view, as he concedes that the distinction between complete and partial excusing conditions is a "crude dichotomy," which "ignore[s] the ever interesting and ever-illuminating varieties of case" (p. 79). This implies that it is possible in principle to partially suspend the reactive attitudes toward both categories of person, and so we are licensed to merely excuse both cultural outsiders and psychopaths. This is consistent with our moral practice, which admits degrees of blame, resentment, disapprobation, and the like.

**6. A Quality of Will Objection**

It has been suggested to me in professional correspondence that an internalist account which emphasizes the quality of the agent's will is preferable to externalism because *quality of will* seems essential to Strawson's philosophy, insofar as Strawson believes that the reactive attitudes must be deployed toward an agent's good or ill will, not his overt behaviour as such. This section is devoted to answering this objection, as well as further clarifying the externalist method and its distinct advantages.

My first objection enlarges upon Sneddon's claim that, "in the case of at least some externalistic competences [including responsibility], *no particular intrinsic properties of an agent may be necessary or sufficient for their realization*" (p. 242, emphasis added.) That is, internalism cannot explain the capacity for moral responsibility. I believe that this is a convincing argument, but Sneddon perhaps does not provide sufficient clarification as to what is wrong with the idea of individually necessary internalistic conditions. I believe that this can be done by comparing moral competence with reading competence. In academic contexts, reading incompetence can be explained, and partially excused, by having dyslexia. As Rita Carter observes (1998), "dyslexia takes many different forms and has many different causes" (p. 154). In some cases, it can be caused by a defect of the insula of the cerebral cortex, preventing this mechanism from firing in unison with the other language areas of the brain. However, an insula defect is only one possible cause of dyslexia: there

are many other etiological possibilities, and most are not scientifically identifiable or diagnosable. Diagnostic testing is usually done on the basis of overt symptoms such as speech delays, letter reversal, and susceptibility to distraction. While functional magnetic resonance imaging (fMRI) tests can highlight insula defects, they are useless for other neuropathologies, and quite impractical for the purpose of general diagnosis. (They are rarely used for any learning disability.) At least for the foreseeable future—and perhaps in perpetuity due to the incredible complexity of the human mind—it is impossible to diagnose dyslexia exclusively on neurobiological grounds. This has implications for moral responsibility. As with reading competence, moral responsibility, i.e., the capacity to respond appropriately to the reactive attitudes, does not seem reducible to neurological or psychological criteria. Even if it is possible in principle to perform this kind of reduction—which, I have said, is rather dubious—defining moral responsibility in concrete psychological or neurobiological terms is presently impossible, and will likely never be practicable is a general strategy. While we might want to postulate some metaphorical, inner defect, we should realize that this approach is symbolic, and that it is no different from, or more informative than, categorizing dyslexia in terms of a single, invariant inner deficit. The diagnostic techniques are still externalistic in nature.

My second argument is that even if internalism were sufficient to account for certain types of excuses, it would not be adequate for every type of excuse. Aside from PUFC, an externalist account appears necessary for explaining judgments of culpable negligence and recklessness. This is evident if we consider transgressions that are both criminally liable and morally blameworthy, which have provoked substantial changes in the context of legal theory and practice. If we recall the landmark case of S. Hundal, it provoked a change in Canadian criminal law whereby the court would no longer require proof of "a subjective mental element of an intention to drive dangerously," but would instead measure responsibility "on the basis of an objective standard without establishing the subjective mental state of the particular accused" (J.E. Bickenback, p. 255). The law, I believe, is ahead of the curve in this regard. Moral philosophy, on close scrutiny, likewise requires an externalist standard for responsibility, to explain our negative reactive attitudes to moral agents who fail to fulfill their objective moral obligations. Indifference, it has been argued, is inadequate for this purpose, because one must consider *why* an agent became indifferent in the first place: if it was due to coercive childhood factors, a further criterion is still needed to explain our reactive attitudes. So at the very least, an externalist criterion is required to exhaust the scope of our reactive attitudes, without leaving gaps.

My final objection relates back to the argument from situationist psychology which has been elaborated at length in this treatise—namely, the epistemological problem. This problem implies that subjective aspects of human psychology are explanatorily effete: as the

Milgram experiment illustrates, situational variables are stronger explanatory and predictive variables than individual character traits. If this is correct, then internalist theories of responsibility, including the quality of will view, are empirically untenable. (For more on this, consult previous chapters, especially chapter 2.)

**7. Concluding remarks**

I do not expect the arguments of this chapter to be dispositive on the question of moral responsibility. One can, of course, still raise reasonable objections to the externalist purview and the externalistic justification of PUFC, though I hope to have responded to the most salient. My primary purpose in writing this chapter was not to silence dissent, but to shift the burden of proof in this debate. Specifically, I hope to have shown that internalists can no longer assume without argument that moral responsibility is an internalistic concept, or that excusing conditions are necessarily and universally reducible to internal, psychological deficits. At the very least, I hope to have shown that it is incumbent upon internalists to acknowledge and defend their (hitherto tacit) methodological assumptions, and argue that standards of moral responsibility are *not* externalistically construed social competences. Furthermore, if they are Strawsonians, they must also show that there is an internalistic interpretation of the reactive attitudes that does not violate Strawson's requirement of interpersonal justification. On this note, I pass the philosophical gauntlet to challengers.

**Chapter 7: The Case of JoJo and Our Pretheoretical Intuitions**

**1. Preliminaries**

I have alluded to the fact that David Faraci and David Shoemaker (2010) object to Susan Wolf's sane deep-self view (SDSV) of moral responsibility, which is supposed to fit better with our pretheoretical intuitions about deprived childhood victims than the plain deep-self view (DSV.) Wolf's account hinges on the intuitiveness of a particular example, which presents JoJo as the son of an evil dictator of a small, undeveloped country, who grows up to adopt his father's sadistic worldview and perpetuate it. Faraci and Shoemaker conduct a survey to test naïve moral reasoners' intuitions on this case. From the study results, they draw three general conclusions: (i) that SDSV is false, (ii) that DSV is adequate, subject to the addition of a 'scalar dimension' which measures degrees of responsibility, and (iii) that JoJo is partially excused on a naïve moral perspective because he is seen as suffering from a rare and peculiar form of moral ignorance, "different in kind from the moral ignorance people usually experience" (Faraci and Shoemaker, p.

327). I dispute these conclusions and present an alternative interpretation of the results, which defends the externalist, Strawsonian account of moral responsibility presented in the last chapter. I then use this approach to show that JoJo's circumstances are inherently extenuating (at least to a significant degree.) Unlike SDSV and DSV, my proposal is capable of accommodating the fact that JoJo's formative circumstances were the *only factor* that made a statistically significant difference to the subjects' moral judgments. This is because it is the only view capable of according *any* intrinsic weight to external, historical factors.

**2. Introduction**

In the *Review of Philosophy and Psychology* (2010), David Faraci and David Shoemaker analyze Susan Wolf's well-known example of JoJo, which is meant to show that the excuse of moral insanity is supported by "our pretheoretical intuitions" (p. 382), in accordance with the sane deep-self view. Faraci and Shoemaker test Wolf's theory by asking laypeople to assess JoJo's blameworthiness across three scenarios that vary his formative circumstances and exposure to moral alternatives, to assess whether these factors make a difference to naïve moral reasoning. The authors ultimately find that moral insanity does not make a statistically significant difference. In their philosophical analysis, they actually draw three separate conclusions: (i) that SDSV is false, (ii) that DSV is basically adequate, subject to the addition of a scalar dimension which measures degrees of moral responsibility, and (iii) that JoJo is partially excused by naïve moral reasoners because he suffers from a peculiar form of moral ignorance, "different in kind from the moral ignorance people usually experience" (p. 327). However, the condition of moral insanity is deemed irrelevant.

In this chapter, I call into question Faraci and Shoemaker's conclusions. First, I contend that the survey results do not support DSV, even with the addition of a scalar component, since the scenarios imply that JoJo's moral ignorance (and consequent character flaws) are embedded at the level of his deep self; thus, JoJo cannot be even partially excused on DSV's evaluative criteria. Secondly, I argue that the authors' explanation of the subjects' judgments is deficient, since it invokes implausibly sophisticated philosophical distinctions, which depart from the text of the scenarios. The correct explanation, I submit, must invoke JoJo's formative circumstances, which were the only variable that patently influenced subjects' thinking. Moreover, even if the excuse of moral ignorance is somehow excusing to unsophisticated reasoners, it is unclear how DSV can accommodate it, since JoJo's moral ignorance is, *ex hypothesi*, ingrained in his deep self.

After raising these objections, I propose an externalist, Strawsonian theory of moral responsibility, which better explains the survey results. This theory accomodates the fact that the only variable that made a significant difference to the subjects' judgments was JoJo's formative

circumstances. It also offers the simplest explanation of the subjects' judgments based on the phrasing of the scenarios, which describes JoJo as evil. This does not mean that Faraci and Shoemaker's analysis is otiose: their argument for the addition of a scalar component is long overdue in moral philosophy. However, it implies that, at the very least, DSV needs to be supplemented by an externalistic condition, to explain the role that situational factors play in moral reasoning; and at worst, in light of other weaknesses, DSV may simply be unsalvageable.

Before embarking on this argument, a brief clarification is in order. In personal correspondence, Faraci has informed me that he and Shoemaker did not mean to suggest that DSV is incompatible with externalistic excusing conditions such as PUFC: perhaps such conditions can render an agent's deep self particularly shallow.[11] But be that as it may, I still cannot see how DSV can be seen to accommodate this type of excuse as *non-derivatively significant*, without some procrustean explanation. In what follows, I will argue that, without some story for how DSV, as an internalistic theory, can be reconciled with inherently *externalistic* excuses (and I myself cannot imagine how such a story would go,) it must be seen as insufficient. I will then build on previous arguments to show that externalism easily accommodates such excuses.

**3. Wolf's Sane Deep-Self View**

We are by now very familiar with Wolf's argument, and the fact that SDSV is defended by appeal to our pretheoretical intuitions and interpersonal consensus. However, we might request empirical evidence to the effect that most people *do* in fact deeply and intuitively value moral sanity as a condition of moral responsibility. Faraci and Shoemaker contend that this cannot be done, because naïve moral reasoners, in point of fact, are not inclined to excuse insanity. At the start of their paper, they relate that, "whenever we introduced the case [of JoJo] to students, they always needed considerable coaching to come to the conclusion Wolf wants" (p. 324). To test Wolf's premise, they conducted a survey on laypeople using three versions of the JoJo scenario, which vary his formative circumstances and exposure to moral alternatives. On the first scenario:

> JoJo is an evil and sadistic dictator of a small, undeveloped country. JoJo does many things as a dictator, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things. When he steps back and asks, "Do I really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life reflects his deepest values and ideals. (p. 325)

[11] For a more fulsome explanation of Faraci and Shoemaker's theory, see "Huck Versus JoJo: Moral Ignorance and The (A)symmetry of Praise and Blame," forthcoming in *Oxford Studies in Experimental Philosophy*.

This scenario is 'the control.' The second and third scenario stipulate that JoJo is "entirely cut off from the outside world" (p. 325), and the third scenario adds the condition that "when he turns 21, JoJo is sent to live in a developed country for a year, and there he becomes aware that other leaders treat their subjects with respect and goodwill because they value the lives and well-being of their subjects. Nevertheless, when he returns to lead his country, JoJo does the same sorts of things his father did…" (p. 326).[12] The authors hypothesize that JoJo will be deemed blameworthy in each of these scenarios, "albeit perhaps less blameworthy than someone like Uday Hussein, and that the degrees of blameworthiness would increase in correspondence to his exposure to relevant moral alternatives" (p. 324). To test this, they ask their subjects to rate JoJo on a scale from 1 ('not at all blameworthy') to 7 ('completely blameworthy,') with 4 being 'somewhat blameworthy' (p. 324).

They obtained the following results: for the control scenario, the mean was 5.7 out of 7; for the second scenario, the mean was 4.77; and for the third scenario, the mean was 4.93, with a mode and median score of 5. The experimenters draw the following conclusions:

(1) Wolf's theory (SDSV) is false because the students deemed all of the JoJos to be more than somewhat blameworthy. (They assigned each JoJo a mean score higher than 4.)

(2) Their hypothesis—that JoJo's access to moral alternatives would explain the subjects' judgments—initially faces certain "complications" (p. 327). Namely, the authors are "genuinely surprised" that JoJo's childhood circumstances "render him less blameworthy (according to our subjects) than the control case seemingly regardless of his moral ignorance" (p. 327). In other words, childhood circumstances seem to be independently morally relevant, regardless of the agent's exposure to moral alternatives. However, the authors offer another explanation which may redeem their initial hypothesis. They argue that while it may seem that exposure to moral alternatives is statistically insignificant, "people could well be judging that mere exposure to moral alternatives is insufficient to dispel moral ignorance, especially for those who have been as thoroughly indoctrinated as JoJo" (p. 327). That is, subjects may be thinking that JoJo3 is particularly susceptible to "confirmation bias," and that this partially excuses him (p. 327). Due to this bias, JoJo2 may be ignorant that there is a demand for goodwill on his part, and so he may be suffering from "a particularly insidious type of ignorance…, a type [that is] is different in kind from the moral ignorance people usually experience" (p. 327). Faraci and Shoemaker posit that this consideration may figure prominently in their subjects' thinking,

[12] Henceforth, the subjects of these scenarios will be referred to as JoJo1, JoJo2, and JoJo3 respectively.

and may have extenuating force for them when they assess JoJo's conduct.

(3) The third conclusion is that the survey results show that DSV is superior to SDVS, conditional upon the addition of a scalar component measuring degrees of blameworthiness. However, this view also faces certain difficulties. The authors concede that it is susceptible to a "Wolfian response," to the effect that, although JoJo2 was deemed somewhat blameworthy, he was still judged significantly less blameworthy than JoJo1, and this must logically be due to "the statement of [his] unfortunate formative circumstances" (p. 327). That is, the formative circumstances must be responsible for the subjects' judgments, as opposed to, say, the inference of especially insidious ignorance, or the postulation of a weak deep self. In response, the authors speculate that, "one might think that [JoJo2] is as blameworthy as he is precisely because he meets the conditions of the DSV, just not as fully as does JoJo1"; and, "at any rate, there may be scalar resources within the DSV itself that could explain both the significant degrees of blameworthiness attached to both JoJos as well as the disparity between them" (p. 327). Consequently, they infer that, on the purview of the subjects, JoJo2 may partially fail to satisfy DSV due to his extreme degree of moral ignorance.

As aforesaid, Faraci has since clarified to me that he and Shoemaker did not intend to exclude unfortunate formative circumstances as potentially extenuating on DSV (perhaps by rendering JoJo's deep self particularly shallow.) However, they propose insidious ignorance as the most likely explanation in their analysis; and they also leave open how formative circumstances are supposed to be excusing on DSV, given that such circumstances do not affect JoJo's present *deep self* in any way, shape, or form. Thus DSV is susceptible to Watson's criticism (1987) that facts about background are merely evidential, not excusing in themselves (p. 274). Therefore, while I take Faraci's point that his interpretation is not meant to rule out formative circumstances, something needs to be said about *how*, i.e., by *what explanatory mechanism*, DSV can accommodate this condition as a first-order excuse. In the remainder of this section, I explain why, without further elaboration, the authors' arguments fail in this regard, and then I offer an *externalist* interpretation which solves this problem.

Returning to (1)-(3), I submit that while these explanations are philosophically interesting and may account for some people's—perhaps some philosophers'—moral intuitions, they do not, in their present articulation, capture the pretheoretical intuitions of naïve moral reasoners. In particular, they do not give sufficient weight to the role that formative circumstances play in the naïve subjects' judgments, in light of the difference in average responses to JoJo1 and JoJo2 (which implies that exposure to moral alternatives was relatively insignificant,) versus the similarity in average responses to JoJo2 and JoJo3 (which suggests that formative circumstances

were significant *regardless of considerations of moral ignorance*—in other words, *inherently* significant.) Faraci and Shoemaker give a somewhat equivocal explanation of their subjects' judgments, acknowledging on the one hand that, "it seems that consideration of JoJo's unfortunate formative circumstances dominates his level of exposure to moral alternatives: these circumstances render him less blameworthy (according to our subjects) than the control case seemingly regardless of his moral ignorance" (p. 327); but on the other hand, they fail to explain how DSV is supposed to accommodate this hypothesis. JoJo is described as wholeheartedly evil, so why would DSV excuse him? While the authors proffer a scalar version of DSV as part of their explanation, it is not, in any obvious way, capable of assigning *any* intrinsic weight to formative circumstances. This is because DSV only takes into account the agent's quality of will, and all three scenarios explicitly state that JoJo *wholeheartedly endorses his actions*. Thus, he is blameworthy on DSV's criteria. This consideration is precisely what precipitates Wolf's criticism of DSV and prompts her to propose SDSV: she recognizes that an additional condition is needed to capture the non-will-related aspect of JoJo's apparent lack of freedom. It is puzzling how adding a scalar dimension to DSV is supposed to solve this problem, when DSV is not able to take historical factors into consideration *to any extent whatsoever*. Adding a scalar dimension might allow us to ascribe proportional blame based on the agent's degree of ill will, but it does accord any weight whatsoever to childhood neglect (except perhaps derivatively, insofar as it precipitates ill will.) But this derivative influence does not accommodate the fact that in the survey, formative circumstances were (seemingly) *intrinsically extenuating* for the subjects. Thus, DSV is inadequate, barring a story that makes sense of how historical factors can affect deep-self-based judgments. In the next section, I present an alternative account that much more easily accommodates historical excuses, based on Strawson's well-known model.

First, however, I should address what I take to be independently mistaken in Faraci and Shoemaker's conclusions, to motivate this alternative perspective. In the first place, if (2) is right and the subjects are excusing JoJo on grounds of *moral ignorance*, it is not clear how DSV is supposed to accommodate this condition (let alone formative circumstances.) This is because if JoJo is morally ignorant, this does not affect the fact that he is a wholehearted sadistic torturer at the level of his deepest self, as the relevant scenarios explicitly state: "When he steps back and asks, 'Do I really want to be this sort of person?' his answer is resoundingly 'Yes,' for this way of life reflects his deepest values and ideals" (Faraci and Shoemaker, p. 325). Secondly, when Faraci and Shoemaker try to explain their subjects' commensurate judgments of JoJo2 and JoJo3, in spite of the latter's exposure to moral alternatives (which genuinely surprised them,) they surmise that the subjects may be seeing the JoJos as sharing "a particularly insidious kind of ignorance" (p. 327), which is categorically different from the type that one typically finds in the general population. Further, this 'insidious ignorance' must be seen as partially exculpating

according to (scalar) DSV. However, it seems to me that the postulated categorical difference is a very sophisticated, fine-grained philosophical distinction to make—essentially requiring differentiation between moral insanity and 'sane' moral ignorance—and I question whether naïve moral reasoners are capable of making this discrimination. The authors' explanation implies that laypeople generally have the ability to understand the difference *between* right and wrong (which accounts for moral sanity,) and "knowledge *of* what is right and wrong" (Faraci and Shoemaker, p. 328), where the latter is conceptually distinct from moral insanity, and yet also not ingrained in the agent's deep self. Moreover, we are supposed to attribute this judgment to the subjects as an explanation of their response to the survey, even though the only textual difference amongst the three scenarios (aside from JoJo3's statistically-insignificant exposure to moral alternatives) was JoJo's formative circumstances. On grounds of simplicity and psychological realism, the formative-circumstances explanation is obviously the most logical construal of the subjects' judgments. Moreover if the subjects are supposed to be blaming JoJo on the basis of his defective quality of will, as DSV requires, it is far from clear that they are capable of distinguishing his (blameworthy) quality of will from his (exculpating) insidious moral ignorance. In standard interpersonal relationships, moral ignorance tends to be *resented as a moral failing*, not forgiven. So we would expect subjects to judge a morally ignorant figure as doubly blameworthy, not excusable. For all these reasons, it is doubtful that DSV manages to capture the subjects' actual moral deliberations.

These criticisms are not conclusive: whether laypeople are actually influenced exclusively by quality-of-will considerations, or by considerations of avoidability, formative circumstances, and historical factors, is ultimately an empirical matter. However, based on the available evidence, it is premature to judge DSV adequate. My arguments are meant to motivate further study into this matter, and different research directions.

In the next section, I present my externalistic interpretation of the results, which will provide an alternative framework for understanding moral reasoning.

**4. Externalism: A Better Explanation**

We have already seen Sneddon's externalist account of Stawsonian responsibility (chapter 6,) so I will not reiterate it here. For present purposes, Sneddon's account is instructive because it lays the groundwork for an alternative interpretation of Faraci and Shoemaker's results. Specifically, it suggests that naïve moral reasoners might be inclined to excuse people on grounds of deprived childhood circumstances, not merely as an evidential factor, but as an inherent, first-order excuse. Previously, I argued that Sneddon's tennis example sheds light on how PUFC can be intrinsically excusing. The same type of analogical example, I believe, can be used to explain why JoJo3's formative circumstances were significant, but not his exposure to moral alternatives at

age 21—a result that defied the experimenters' hypothesis. To this end, I want to use an example that depends particularly heavily on early childhood education—namely, the acquisition of speech. Scientists have substantiated that children who do not learn language within the first decade of life never acquire the ability to form grammatical sentences (Carter 1999, p. 146). This was confirmed when, tragically, a 13-year-old child was discovered in a solitary, squalid, featureless room in Los Angeles in the 1970s. Due to extreme social isolation, the girl had never learned how to speak. When scientists tried to teach her English, she was unable to achieve grammatical competence: after considerable instruction, she could still only give one-word answers to questions (Carter, p. 156). She never achieved linguistic competence. A similar example involves people who learn a second language as adults; they almost never manage to lose their accent, even if they take great pains to improve their elocution. This is because certain phonemes do not translate across languages: for example, "Japanese adults cannot reproduce the English 'l' and 'r'… because these are sounds not heard in Japanese" (Carter, p. 156), and Anglophone adults cannot reproduce certain Japanese linguistic units. This suggests that interpersonal deficits acquired in childhood (through social isolation, abuse, or lack of education) tend to persist into adulthood, regardless of how much effort one makes to offset them. Hence, if moral competence is like linguistic competence, then just as we cannot reasonably expect a feral child to learn grammar, or a native Japanese speaker to reproduce the English 'l' and 'r' phonemes, we cannot reasonably expect deprived childhood victims to acquire moral competence, even if they are exposed to moral alternatives. This is because, in moral and non-moral development alike, rehabilitative learning strategies do not seem to eliminate deficits acquired in childhood, as scientists discovered when they tried to teach grammar to their neglected subject. *A fortiori*, mere *exposure* to alternatives (without active teaching) is unlikely to make a significant impact. Of course, this claim is case sensitive, with the degree of excusing depending on the degree of deprivation. While some people can be only mildly excused, others may be exempted in light of their personal history. This helps to explain why subjects were inclined to (partially) excuse JoJo3 in spite of his exposure to moral alternatives as an adult.

Another important implication of Sneddon's view, recall, is that externalistically-construed deficits can be acknowledged and regarded as excusing even if underlying psychological, cognitive, or neurobiological deficits resist diagnosis or identification. This was illustrated by comparison with dyslexia. As we saw, dyslexia is typically diagnosed on the basis of questionnaires about the child's development, education, and medical history, but almost never through fMRI. Further, although certain forms of dyslexia are identifiable in neurological terms—for example, those due to defects of the insula of the cerebal cortex—most are not identifiable on this basis, and may never, for all we know, be reducible to neurological mechanisms. However, this non-reducibility does not prevent 'dyslexia,' referring to a cluster of behavioural deficits, from constituting a legitimate excusing condition for reading incompetence. We would not deny a

student accommodations for lack of neurological proof. Moral competence, similarly, is primarily displayed at the behaviour level, through an agent's ability (or inability) to respond to the reactive attitudes of others. By parity of reasoning, we ought to acknowledge interpersonal deficits displayed in over behaviour as legitimate excusing conditions, unless we have reason to suspect malingering or other forms of deceit. Thus, like developmental disabilities, moral deficits are susceptible to externalistic, symptomatic explanation.

A final point, which bears more directly on *moral capacities,* is that longitudinal research in social psychology may provide evidence about moral capacities that could implicitly influence our moral judgments. In 'Unequal Childhoods' (2011), a longitudinal study on childhood development, Annette Lareau followed 88 children from different socioeconomic backgrounds for several years, and performed follow-up interviews ten years later. She found that *concerted cultivation* (i.e., the practice of encouraging children to develop their social talents) directly contributed to the acquisition of advanced language use, reasoning, negotiation, organizational skills, and other social competences. Furthermore, rather than diminishing over time, the differential advantages acquired in childhood tended to accrue: "Young adults from middle-class families were more likely to graduate from high school, apply to four-year colleges, gain admission, and enroll," and social gains "were more noticeable by the time the youth had become adolescents" (Lareau, p. 278). This finding corroborates the idea that adults may not be able to compensate for childhood deficits, no matter how hard they try. This finding, I believe, may be available to observation, and thus may give rise to an implicit commonsense intuition, held by perspicuous people, to the effect that neglected childhood victims are permanently disadvantaged in ways that people from privileged backgrounds are not, and therefore are not fully responsible for their deficits in social competence. If I am right, this would explain the fact that Faraci and Shoemaker's subjects were inclined to excuse JoJo2 and JoJo3 roughly equally, even though JoJo3 was later exposed to moral alternatives. In line with Lareau's findings, subjects may be thinking that JoJo3 is permanently morally disabled on account of his deprived childhood, and longitudinally incapable of acting morally. In addition, I suspect that subjects may be thinking this regardless of considerations of JoJo's psychology—that is, they may not particularly care whether there is an underlying incapacity such as 'moral insanity,' a lack of 'second-order volitions' or 'evaluative motives.' The fact of JoJo's deprived childhood in and of itself may suffice to excuse his behaviour. (This resembles my earlier description of entrapment, in which an authority figure's criminal inducements partially excuse the recipient, even if the latter gladly accepts the inducements, because society has a stake in discouraging professional misconduct. Thus, entrapment inherently excuses, so to speak). If this is right, then formative circumstances are inherently excusing to naïve moral reasoners, regardless of whether they can be inscribed onto internal processes. This lends credence to the view that an externalist condition of moral responsibility is needed, either as a supplement to an internalistic view an independent theory. In

light of the epistemological problem, it may be reasonable to eschew all internalistic approaches.

One might object that a Strawsonian account should excuse JoJo to a greater extent than Faraci and Shoemaker's subjects did, i.e., the score should be lower than 4.77 to 4.93 out of 7. It is unclear whether this is so, since the degree of influence of each excusing factor is not indicated in Strawson's theory. The main advantage of the externalist-Strawsonian view is that it explains why formative circumstances would make any intuitive difference at all, since internalist accounts are impotent in this regard. The degree of excusing force is something that would need to be determined via investigation. In any case, externalism has the advantage of acknowledging the minimal moral relevance of situational excuses.

**5. Concluding Remarks**

These considerations do not silence criticism, and I have acknowledged that more research is needed, but I hope to have provided a new avenue for investigation. Faraci seems to be of the opinion that if internalism is plausible, it must be a supplement to DSV; but in light of the epistemological problem, DSV seems empirically untenable. Moreover, DSV still would not be able to explain why naïve reasoners seem to excuse JoJo even though he is wholeheartedly evil. Perhaps Faraci will explain these anomalies in subsequent scholarship, but for now, I think that the externalist Strawsonian account is preferable, and quite self-sufficient. Where I do agree with Faraci and Shoemaker is that moral philosophy would benefit from the acknowledgment of a scalar component, which would allow us to recognize degrees of responsibility. This would help to address the problem that I confronted in chapter 5, i.e., that moral philosophers too readily exclude people with mental illness from the category of personhood, thereby extinguishing their basic rights and freedoms. This is far from an academic problem, as psychiatric patients are currently suffering the indignity of forced hospitalization and community treatment orders (CTOs) within my own community of Toronto at an alarming rate. This phenomenon is documented by Erick Fabris in his book, 'Tranquil Prisons: Chemical Incarceration under Community Treatment Orders' (2011). I hope that my research can be used to rectify this problem, as well as remediating the equally pernicious injustice of denying children basic resources that contribute to moral development. Doing so is not only unconscionable from a rights perspective, but also likely to cause deleterious effects in the form of sociopathic behaviour, for which no one is more responsible than those who failed to ensure a fair distribution of moral resources. This is one of the key lessons of developmental psychology, which all should heed.

**References**


Anderson, E. 1993. What is the point of equality*? Ethics* 109, pp. 287-337.

Archard, D. 1999. The *mens rea* of rape: Reasonableness and culpable mistakes. In Keith Burgess-Jackson (ed.), *A 'most detestable crime': New philosophical essays on rape*. New York: Oxford University Press.

Arendt, H. 1963. *Eichmann in Jerusalem: The banality of evil*. New York: Penguin Books.

Anderson, E. 1995. Feminist epistemology: An interpretation and a defense. *Hypatia* 10(3), pp. 50–84.

Baumeister, R.F., and Vohs, K.D. 2004. Four roots of evil. In A.G. Miller (ed.), *The social psychology of good and evil* (pp. 85–101). New York: The Guilford Press.

Bentall, R.P. 2004. *Madness explained: Psychosis and human nature*. London: Penguin Books.

Bickenbach, J. E. (Ed.) 2007. *Philosophy of law 4th edn*. Peterborough, ON: Broadview Press.

Brown, S.M. 1950. Does ought imply can? *Ethics* 60(4), pp. 275-284.

Cane, P. 2002. Responsibility in law and morality. Oregon: Hart Publishing.

Carter, R. 1998. *Mapping the mind*. London, England: University of California Press.

Cartwright, W. 2006. Reasons and selves: Two accounts of responsibility in theory and practice. *Philosophy, psychiatry & psychology* 13(2), pp. 143-155.

Ciurria, M. Answering the situationist challenge: A defense of virtue ethics as preferable to other theories. *Dialogue*, forthcoming 2014.

Code, L. 1991. *What can she know?* New York: Cornell University Press.

Cohen, J.M., and Blesdoe, C. 2002. Immigrants, agency, and allegiance: Some notes from anthropology and from law. In R. Schweder et al. (eds.), *Engaging cultural differences: The multicultural challenge in criminal democracies*. (pp. 99-127). New York: Russel Sage.

Copp, D. 2008. "Ought" implies "can" and the derivation of the principle of alternate possibilities. *Analysis*, pp. 67-75.

Culver, Keith C. (ed.). 2008. *Readings in the philosophy of law*, 2nd edn. Peterborough, ON: Broadview Press.

Cummiskey, D. 1996. *Kantian consequentialism*. Oxford: Oxford University Press.

Cushman, F., Young, L., and Greene, J. D. 2010. Multi-system moral psychology. In J. Doris (ed.), *The moral psychology handbook* (pp. 47-71). Oxford: Oxford University Press.

Daniels, N. 1979. Wide reflective equilibrium and theory acceptance in ethics. *Journal of philosophy* 76(5), pp. 256-282.

Darley, J.M. and Batson, C.D. 1973. From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of personality and social psychology* 27, pp. 100-108.

Doris, J. M. 2007. Out of character: on the psychology of excuses in the criminal law. In H. Lafolette (ed.), *Ethics in practice*, 3rd edn. (pp. 519–530). Malden: Blackwell Publishing.

_____ 2002. *Lack of character: Personality and moral behavior*. New York: Cambridge University Press.

_____ 1998. Situations and virtue ethics. *Nous*, 32(4), pp. 504–530.

Duff, R.A. 1993. Choice, character and criminal liability. *Law and Philosophy* 12, pp. 345-383.

Duntley, J.D., and Buss, D.M. 2004. The evolution of evil. In A. G. Miller (ed.), *The social psychology of good and evil* (pp. 102–125). New York: The Guilford Press.

Dworkin, R.M. 1977. *Taking rights seriously*. Cambridge: Harvard University Press.

Ebert, T. 1996. *Lucid feminism and after: Postmodernism, desire, and labor in late capitalism.* Michigan: University of Michigan.

Edmonds, D. 2014. *Would you kill the fat man?* Princeton, NJ: Princeton University Press.

Ekman, P. & O'Sullivan, M. 1991. Who can catch a liar? *American psychologist* 46 (9), pp. 913-920.

Fabris, Erick. 2011. *Tranquil prisons: Chemical incarceration under community treatment orders*. Toronto: University of Toronto Press.

Faraci, D. and Shoemaker, D. 2010. Insanity, deep selves, and moral responsibility: The case of JoJo. *Review of philosophy and psychology* 1(3), pp. 319-332.

Fascher, J.M. 1988. Responsiveness and moral responsibility. In F. Shoeman (ed.), *Responsibility, character, and the emotions* (pp. 81-106). Cambridge, Cambridge University Press.

Finnis, J. 1980. *Natural law and natural rights*. Oxford: Oxford University Press.

Fischer, J.M. 2002. Frankfurt-style compatibilism. In G. Watson (ed.), *Free will 2nd edn.* (pp. 190-211). Oxford, Oxford University Press, 2003.

Flanagan, O. W. 1991. *Varieties of moral personality*. Cambridge: Harvard University Press.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 67, pp. 5-20.

Fricker, M. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford: Oxford University Press.

Gazzaniga, M.S. 2005. *The ethical brain: The science of our moral dilemmas*. New York: Harper Collins.

_____ 1992. *Nature's mind: The biological roots of thinking, emotions, sexuality, language, and intelligence.* Harmondsworth: Penguin.

Gazzaniga, M.S., LeDoux, J.E., and Wilson, D.H. 1977. Language praxis and the right hemisphere: Clues to some mechanisms of consciousness. *Neurology* 27, pp. 1144–1147.

Gilbert, P. 2010. *Compassion focused therapy*. London: Routledge.

Gimenez, M. 1990. *Work without wages: Comparative studies of domestic labor and self-employment*. Suny Press.

Gino, F. 2013. *Sidetracked: Why our decisions get derailed and how we can stick to the plan*. Boston: Harvard Business School Publishing.

Goodman, N. 1953. The New Riddle of Induction. In his *Fact, fiction, and forecast*. Harvard: Harvard University Press.

Graham, G. 2010. *The disordered mind: An introduction to philosophy of mind and mental illness.* Oxon: Routledge.

Granhag, P.A. and Strömwall, L.A. (eds.) 2004. *The detection of deception in forensic contexts*. Cambridge, UK: Cambridge University Press.

Greenberg, G. 2013. *The book of woe: The DSM and the unmaking of psychiatry*. New York: Blue Rider Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. 2004.The neural bases of cognitive conflict and control in moral judgment, *Neuron* 44, pp. 389 – 400.

Haidt, J., & Hersh, M. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology* 31, pp. 191–221.

Harman, G. 2000. The non-existence of character traits. *Proceedings of the Aristotelian society new series 100*, pp. 223-226.

_____ 1999. Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian society 99*, pp. 315**–**331.

_____ 1975. Moral relativism defended. *Philosophical review* 84(1), pp. 3–22.

Hart, H.L.A. 1994. *The concept of law 2nd edn*. Oxford: Clarendon Press.

_____ 1953. The aims of the criminal law. *Law and cotemporary problems* 23(3), pp. 401- 441.

Hartshorne,H., and May, M.A. 1928. *Studies in the nature of character, vol. 1*. New York: Macmillan.

Harvard Law Review Association (HLRA) (1986). The cultural defense in the criminal law. *Harvard Law Review* 99(6), pp. 1293-1311.

Hirstein, W. 2005. Brain fiction and the riddle of confabulation. Cambridge: MIT.

Holmes, O.W. 1881. *The common law*. In M.D. Howe (ed.), Cambridge, Mass: Harvard University Press, 1963.

Hooker, B. 2000. Ideal code, real world. Oxford: Clarendon Press.

Hundal v R [1993] 1 S.C.R. 867.

Hurley, P. 2009. *Beyond consequentialism*. Oxford: Oxford University Press.

Hursthouse, R. 1999. *On virtue ethics*. Oxford: Oxford University Press.

Isen, A.M. and Levin, H. 1972. Effect of feeling good on helping: Cookies and kindness. *Journal of personality and social psychology* 21, pp. 384-388.

Kading, D. 1954. Does 'ought' imply 'can'? *Philosophical studies* 5(1), pp. 11-15.

Kahneman, D. 2013. *Thinking, fast and slow*. Anchor Canada.

Kamtekar, R. 2004. Situationism and virtue ethics on the content of our character. *Ethics* 114, pp. 458–491.

Kant, I. 1997. *Groundwork of the metaphysics of morals*. Trans. M. Gregor. Cambridge: Cambridge University Press.

Kelly, T. and McGrath, S. 2010. Is reflective equilibrium enough? *Philosophical perspectives* 24, pp. 325-359.

Kennett, J. 2007. Mental disorder, moral agency, and the self. In B. Steinbock (ed.), *The Oxford handbook of bioethics* (pp. 90-113). New York, NY: Oxford University Press.

Lareau, A. 2011. *Unequal childhoods: class, race, and family life,* 2nd ed. Berkeley and Los Angeles: University of California Press.

Lavallee, R. V. (1990). 1 S. C. R. 852. In J.E. Bickenback (ed.), *Canadian cases in the philosophy of law*, 2nd edn. (pp. 274–280). Toronto: Broadview Press, 2007.

Lemos, N. Epistemology and ethics. 2002. In B. Moser (ed.),The Oxford handbook of epistemology (pp. 479-512), Oxford: Oxford University Press.

Libet, B. (1992). The neural time-factor in perception, volition, and free will. Revue de metaphysique et de morale 97(2), pp. 255–272.

Longino, H. 2002. *The fate of knowledge*. New Jersey: Princeton University Press.

MacCloed, M. Charles Whitman: The Texas bell tower sniper. <http://www.crimelibrary.com/notorious_ murders/mass/whitman/index_1.html> Accessed 10 July 2014.

Mandel, D. R. (1998). The obedience alibi: Milgrim's account of the Holocaust reconsidered. Analyze & Kritik 20, 74–94.

Mann, V.S. & Bull, R. 2004. Detecting true lies: Police officers' ability to detect suspects' lies. *Journal of applied psychology* 89, pp. 137-149.

Marquis, D. 2005. What's wrong with adultery? In D. Boonin & Graham Odie (eds.), *What's wrong 2nd edn*. New York: Oxford University Press.

McDowell, J. 1979. Virtue and reason. *Monist* 62, pp. 331–50.

McKenna, M. S. 1998. The limits of evil and the role of moral address: A defense of Strawsonian compatibilism. *Ethics* 2(2), pp. 123–142.

Mendez, M. F., Anderson, E., & Shapria, J. S. (2005). An investigation of moral judgment in frontotemporal dementia. *Cognitive and behavioral neurology* 18(4), pp. 193 – 197.

Merritt, M. 2000. Virtue ethics and situationist personality psychology. *Ethical theory and moral practice* 3(4), pp. 365-383.

Milgram, S. 1974. *Obedience to authority*. New York: Harper and Row.

_____. 1963. Behavioral study of obedience. *Journal of abnormal and social psychology* 67, pp. 371-378.

Miller, G. M., et al. 2002. Explaining the Holocaust: Does social psychology exonerate the perpetrators? In L.S. Newman & R. Erber (eds.), *Understanding genocide: The social psychology of the Holocaust* (pp. 301–325). Cary: Oxford University Press.

Minghella, A. 2000. *The talented Mr. Ripley—Based on Patricia Highsmith's novel*. London: Methuen.

Moore, M.S. 1997. *Placing blame: A general theory of criminal law*. Oxford: Oxford University Press.

Moran, M. 2003. *Rethinking the reasonable person: An egalitarian reconstruction of the objective standard.* Oxford: Oxford University Press.

Morse, S. 2011. Lost in Translation?: An Essay on Law and Neuroscience. In M. Freeman (ed.), *Law and neuroscience, current legal issues vol. 13* (pp. 529-562). New York: Oxford University Press.

Murphy, D. 2006. *Psychiatry in the scientific image*. Cambridge: MIT Press.

Nelkin, D.K. 2011. *Making sense of freedom and responsibility*. Oxford: Oxford University Press.

Newcomb, T.M. 1929. *The consistency of certain extrovert-introvert behavior patterns in 5 problem boys*. New York: Columbia University, Teachers College, Bureau of Publications.

Nichols, S. 2005. *Sentimental rules: On the natural foundations of moral judgment.* Oxford: Oxford University Press.

Nichols, S. and Knobe, J. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs* 41 (4), pp. 663–685

Parfit, D. 2008. *On what matters*, unpublished manuscript.

Patten, S.C. 1977. Milgram's shocking experiments. *Philosophy* 52(202), pp. 425-440.

People v Kimura [1985] No. A-091133

Pereboom, D. 2006. *Living without free will*. New York: Cambridge University Press.

Perring, C. 2011. Bridging the gap between philosophers of mind and brain researchers: The example of addiction. *Mens sana monographs,* Jan-Dec 9(1), pp. 193–201.

_____. 2009. The place of moral responsibility and mental illness. *The American journal of bioethics-neuroscience* 9(9), pp. 32-33.

Pickard, H. 2011. Responsibility without blame: Empathy and the effective treatment of personality disorder. *Philosophy, psychiatry & psychology* 18(3), pp. 209-223.

Prinz, J., and Nichols, S. 2010. Moral emotions. In J. Doris (ed.), *The moral psychology handbook* (pp. 111-146). Oxford: Oxford University Press.

Rawls, J. 1971. *A theory of Justice*. Cambridge, MA. Harvard University Press.

R v Lavallee [1990] 1 S.C.R. 852.

Rosen, G. 2003. Culpability and ignorance. *Proceedings of the Aristotelian society new series,* pp. 61-84.

Ross, L., and Nisbett, R.E. 1991. *The person and the situation*. Philadelphia: Temple

University Press.

Salzinger, S., Feldman, R.S., and Muriel, H., and Rosario, M. 1993. The effects of physical abuse and children's social relationships. Child development 64(1), pp. 169-187.

Scanlon, T.M. 1986. The significance of choice. *The Tanner lectures on human values* vol. 8 (Salt Lake City: University of Utah Press).

_____. 1998. *What we owe to each other*. Cambridge, Mass: Belknap Press.

_____. 2008. *Moral dimensions: Permissibility, meaning, blame*. Cambridge, Mass: Belknap Press.

Sears, D. O. 1994. Ideological bias in political psychology: The view from scientific hell. *Political psychology* 15(3), pp. 547-56.

Sharot, T. 2011. *The optimism bias*. New York: Knopf Canada.

Shoeman, F. 1987. Statistical norms and moral attributions. In F. Shoeman (ed.), *Responsibility, character, and the emotions* (pp. 287-315). Cambridge, Cambridge University Press.

Shroeder, T. 2005. Moral responsibility and Tourette syndrome. *Philosophy and phenomenological research* 71(1), pp. 106-123.

Sinclair, J. 1992. Bridging the gaps: an inside-out view of autism (or, Do you know what I don't know?). In Schopler and Mesibov (eds.), *High functioning individuals with autism* (pp. 294-302). New York & London: Plenum Press.

Sinnott-Armstrong, W., Young, L., and Cushman, Fiery. 2010. Moral intuitions. In J. Doris (ed.), *The moral psychology handbook* (pp. 246-273). Oxford: Oxford University Press.

Slote, M. 1982. Is virtue possible? *Analysis,* 42(2), pp. 70-76.

Smart, J.J.C. 1961. Free will, praise and blame. In G. Watson (ed.), *Free will 2nd edn*. (pp. 291-306). Cambridge: Cambridge University Press.

Smith, B. 2006. Reasons, responsibility, and fiction. *Philosophy, psychiatry & psychology* 13(2), pp. 161-166.

Smith, M. 1995. Internal reasons. *Philosophy and phenomenological research* 55(1), pp. 109-131.

Sneddon, A. 2005. Moral responsibility: The difference of Strawson, and the difference it should make. *Ethical theory and moral practice* 8, pp. 239-264.

_____ 2006. *Action and responsibility*. Netherlands: Springer.

Sosa, E. 2009. Situations against virtues: The situationist attack on virtue theory. In M. Chrysostomos (ed.), *Philosophy of the social sciences: Philosophical theory and scientific practice* (pp. 274-290). Cambridge: Cambridge University Press.

Sperry, R. W. 1961. Cerebral organization and behavior: the split brain behaves in many respects like two separate brains, providing new research possibilities. *Science* 133, pp. 1749–1758.

Staub, E. 2004. Basic human needs, altruism, and aggression. In A.G. Miller (ed.), *The social*

*psychology of good and evil* (pp. 51–84). New York: The Guilford Press.

Stephens, L.G, and Graham, G. 2009. An addictive lesson: A case study in psychiatry as cognitive neuroscience. In M. Broome and L. Bortolotti (eds.), *Psychiatry as cognitive neuroscience*, pp. 203-220. Oxford: Oxford University Press

Stitch, S. 1993. *The fragmentation of reason*. Cambridge, Mass: MIT.

Strawson, G. 1994. The Impossibility of moral responsibility. *Philosophical studies* 75(1/2), pp. 5-24.

Strawson, P. F. 1963. Freedom and resentment. In G. Watson (ed.), *Free will 2nd edn*. (pp. 72–93). Oxford: Oxford University Press, 2003.

Street, S. 2006. A Darwinian dilemma for realist theories of value. *Philosophical studies* 127, pp. 109-166.

Tadros, V. 2011. *The ends of harm: The moral foundations of criminal law*. Oxford: Oxford University Press.

Taylor, C. 1976. Responsibility for self. In A. O. Rorty (ed.), *The identities of persons* (pp. 281–299). Berkeley, Los Angeles: University of California Press.

Waldmann, M. R. and Dieterich, J. H. 2007. Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological science* 18(3), pp. 247–253.

Warren, M.A. 1973. On the moral and legal status of abortion. In H. Lafollette (ed.), *Ethics in practice 3rd edn.* (pp. 126-136). Malden, MA: Blackwell Publishing.

Watson, G. 1987. Responsibility and the limits of evil: Variations on a Strawsonian theme. In F. Shoeman (ed.), *Responsibility, character, and the emotions* (pp. 256-286). Cambridge: Cambridge University Press.

_____ 1975. Free agency. *Journal of philosophy* 72, pp. 205-220.

West, R., Meserve, J., and Stanovich, K. 2012. Cognitive sophistication does not attenuate the bias blind Spot, *Journal of personality and social psychology* 103(3), pp. 1-15.

Wiley, J. S. 1999. Not guilty by reason of blamelessness: culpability in federal criminal interpretation. *Virginia law review* 85(6), pp. 1021-1162.

Williams, G. 2003. Blame and responsibility. *Ethical theory and moral practice* 6(4), pp. 427- 445.

Wolf, S. 1987. Sanity and the metaphysics of responsibility. In G. Watson (ed.), *Free will* (pp. 372-388). Oxford: Oxford University Press.

Zamzow, J. and Nichols, S. 2009. Variations in ethical intuitions. .*Philosophical issues* 19: pp. 368-388.

Zimbardo, P. 2007. *The Lucifer effect: understanding how good people turn evil*. New York: Random House.