UNIVERSITY OF CALIFORNIA
RIVERSIDE

An Animal That is Permitted to Promise: Nietzsche's Way of Naturalizing Responsibility

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Philosophy

by

Avery Jacob Snelson

June 2020

Dissertation Committee:
Dr. Maudemarie Clark, Chairperson
Dr. Pamela Hieronymi
Dr. Coleen Macnamara
Dr. Michael Nelson
Dr. Eric Schwitzgebel

The Dissertation of Avery Jacob Snelson is approved:

_____

_____

_____

_____

_____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

I'd like to thank all of the members of my committee. Thank you for the obvious, for your advice, insights, and comments over the years, for your time, patience, and effort. But thank you also for the not-so-obvious.

Pamela, thank you for your enthusiasm and encouragement, the acuity, economy, and precision of your thoughts, and for catching most of my grammatical mistakes. Had you written this, I have no doubt it would have been half as long, twice as good, and one-hundred times more agreeable to a paralegal.

Eric, much the same applies. Thank you for your timely and encouraging feedback, for reading the longest, messiest, and most impossible drafts I sent to anyone. (No one deserves that kind of punishment.) You're a Confucian saint. I am so fortunate to have had you on my committee.

Coleen, thanks for believing in my Strawsonian reading of Nietzsche in a moment of vulnerability. Not only did you tell me I did not have Strawson wrong, you told me that my Nietzsche was more Strawsonian than Strawson—what a compliment! (You will see this made it in the Intro.) Thank you for your sincerest efforts to assist me in organizing my thoughts and my chapters. Everything I said about Pamela and grammar applies to you and structure.

Michael, six years ago you said that you "liked my work." I had no idea then, and still have no idea today, what you were referring to. But I have always remembered it and your words gave me the courage to invite you onto my proposition committee. Thank you for your feedback way back when—when Dave, Meredith, Zac, Taylor, and I would meet in the library, and for some reason you were willing to read our work. You are the most ecumenical philosopher I've ever met.

Maude, what can one say? I feel like I should quote something from *Schopenhauer as Educator*, but I'm sure one of your other students will do so at some point, and better. So I will do the more Nietzschean thing and quote myself, something I said in a letter I wrote on your behalf some time ago: "Professor Clark's acumen as a great philosopher is perhaps most evident in the way she reads and dissects a text in class: carefully, slowly, critically, ruminating over a particular word choice or phrase, consulting the original German, being careful to situate the central claim or argument within its surrounding context. She does this sort of thing as a matter of *routine* in her seminars, each time conveying to her students, by deed if not in word, *this is how you do philosophy* … she has imparted to me the highest standard of academic rigor, and for that and much else I am deeply indebted to her."

Most of the material from Chapter 3 has been reprinted from an article of the same name in the *Journal of Nietzsche Studies* (Volume 50, Autumn 2019).

To My Bear and Our Cubs

And

To God …

May He Rest in Peace (GS 125)

ABSTRACT OF THE DISSERTATION

An Animal That is Permitted to Promise: Nietzsche's Way of Naturalizing Responsibility

by

Avery Jacob Snelson

Doctor of Philosophy, Graduate Program in Philosophy
University of California, Riverside, June 2020
Dr. Maudemarie Clark, Chairperson

Abstract: I argue that the second essay of Nietzsche's *Genealogy of Morality* (GM II) provides an account of morally responsible agency, rooted in our susceptibility to the feeling of guilt. Specifically, I argue that GM II's naturalistic and developmental account of the origins of conscience, bad conscience, and guilt explains why human beings are, in general, appropriate targets of the moral reactive attitudes. I motivate this reading by appealing to P.F. Strawson's naturalistic analysis of responsibility in "Freedom and Resentment," showing that GM II likewise analyzes responsibility in terms of the *practice* of holding oneself and others responsible. This practice is constituted by what Nietzsche calls the "reactive affects," and I argue that these attitudes not only legitimize blame, but also trust ("permitted promising"), in connection to Nietzsche's positive ideal of responsibility, the "sovereign individual."

CONTENTS

Introduction

I. Free Will and Moral Agency

Friedrich Nietzsche (1844-1900) is perhaps best known for his self-proclaimed "immoralism." As he intends it, immoralism is not the promotion of immoral actions (D 103), nor is it simply the thesis that morality lacks legitimacy or authority over us, a position which one might label "moral skepticism." Immoralism goes beyond moral skepticism by arguing that morality is actually *bad* or *harmful* for us, and thus that we ought to be "immoralists," critics or detractors of morality. In other words, immoralism is a kind of evaluative critique, a critique of morality from an alternative moral perspective. As Nietzsche states in the *Genealogy*, reflecting upon the previous works in which he criticized morality, "The issue for me was the *value* of morality" (GM P: 5), more precisely the "*value* of compassion and of the morality of compassion" (GM P: 6). "We need a *critique* of moral values," he goes on state, "*for once the value of these values must itself be called into question*—and for this we need a knowledge of the conditions and circumstances out of which they have grown, under which they have developed and shifted" (GM P: 6).

My focus in this dissertation will be on a set of related concerns surrounding some of these "conditions and circumstances," specifically Nietzsche's naturalistic account of the development of conscience, guilt, and responsible agency in humans, the practices that were integral to their development on his analysis—punishment and blame—and, finally, the notion of free will, which has traditionally been relied upon to justify them. It is no secret that a significant component of Nietzsche's "immoralism" involves a rejection of the metaphysical presuppositions surrounding human agency that he believes support this "morality of compassion." Chief among these is the belief in libertarian free will (BGE 19, 21) and a corresponding notion of blame that has its roots

in *ressentiment* (GM I: 13, TI VI: 7). According to one prominent interpretation of Nietzsche's immoralism, defended by Brian Leiter (2015), Nietzsche rejects not just libertarian free will and ascriptions of responsibility rooted in *ressentiment*, but *all* ascriptions of responsibility, because he argues that Nietzsche is a "hard-determinist." According to Leiter, agents must perform their actions freely or autonomously to be responsible for their actions. Calling this the "Autonomy Condition," he argues that "Nietzsche's theory of agency involves a sustained attack on the Autonomy Condition, hence on the idea that agents are morally responsible for what they do" (2015: 69). Elsewhere Leiter (2011, 2019b) argues that Nietzsche also rejects the affect of guilt on these same grounds.

Of course, a lot depends—indeed *everything* depends—on what Nietzsche and Leiter mean by "free will," "autonomy," "guilt," and "responsibility," and the relation these bear to human agency more broadly. My aim in this dissertation is to show that there is a type of agency unique to human beings on Nietzsche's analysis, a kind that distinguishes us from other kinds of creatures, and that it is accurate to characterize this as "responsible agency" or "moral agency." Throughout I will use these phrases interchangeably (if one is a responsible agent, then she is a moral agent and vice versa), because each is rooted in our susceptibility to the feeling of guilt on Nietzsche's analysis. Leiter denies this; he denies that we are responsible or moral agents. He does so not only because he believes Nietzsche's theory of agency shows the Autonomy Condition to be false, but also because, as his reliance on the Autonomy Condition reveals, he takes moral agency to be equivalent to the question of whether an agent possesses free will. As he claims, "distinctively *moral* agency … is always in the modern tradition understood to be *free* agency" (Leiter 2019a: 1).

Leiter is *mostly* right. With few exceptions, the "modern tradition" in philosophy has equated "free agency" with "moral agency." Indeed, for much of his career Nietzsche was among the "modern tradition" in equating these ideas, but by the end of his career he was among the few exceptions who did *not*, or at least that is what I will argue. Another exception was P.F. Strawson, whose immensely influential "Freedom and Resentment" (1962) inspired what one might characterize as a *paradigm shift* in the way we think about moral responsibility. In this dissertation I will argue that Nietzsche shared certain tenets central to Strawson's alternative approach. The basis of my reading comes from the second essay of Nietzsche's *Genealogy of Morality* (GM II), wherein he provides a genealogy of conscience, guilt, and responsibility. He even declares GM II to be "the long history of the origins of *responsibility*" (GM II: 2). Maudemarie Clark was the first to observe that this story "fits nicely into a Strawsonian framework" (2015: 90). This dissertation will provide further evidence of her observation.

Central to Strawson's approach was an emphasis on the moral emotions, which he called the "reactive attitudes." He understood these to be responses to an expectation or "demand [for] some degree of goodwill or regard" (1962: 76) on the part of others toward ourselves, *and* on the part of ourselves toward others. He argued that our responsibility practices, principally blame, depended on these attitudes, and, more controversially, that the legitimacy of these attitudes and practices could not be undermined by a general thesis of causal determinism. Strawson's analysis was importantly different from that of the vast majority of philosophers before him. Though his essay was framed in terms of the traditional debate surrounding free will and determinism, as an attempt to "reconcile" the views of libertarians and consequentialist-minded compatibilists, the picture of responsibility he ended up presenting had very

little in common with philosophers before him. Whereas they analyzed responsibility primarily from the standpoint of *metaphysics*, through a conceptual analysis of free will and its relationship to determinism, Strawson's approach was distinctly *naturalistic* in that it instead privileged these attitudes and practices, assuming morality to be a natural phenomenon that evolved along with and was responsive to these attitudes and practices in essential ways. According to this alternative perspective, the Autonomy Condition is not the be-all and end-all of what it means to be responsible.

If the position I attribute to Nietzsche in this dissertation is correct, he agrees with Strawson about the relative unimportance of the Autonomy Condition, because he shares certain beliefs in common with Strawson's naturalistic way of thinking about moral responsibility. Those beliefs are as follows:

1. Our idea of responsibility is ineluctably informed by our moral practices, in particular the practices of blame and punishment;

2. Those moral practices are *constituted* by our moral expectations and their corresponding attitudes, in the sense that these are causal and rational presuppositions of those practices (Strawson calls these the "reactive attitudes," Nietzsche calls them the "reactive affects");

3. Finally, and consequently, an analysis of free will is unnecessary to the issue of *exemption*, the issue of whether one is a moral or responsible *agent*, or one toward whom these expectations and attitudes might be fairly and legitimately applied and felt, because being an apt target of these expectations and attitudes is instead a matter of *moral competence*.[1]

[1] All of these assumptions are Strawsonian *in spirit*, though I am not trying to give an exact and faithful rendering of Strawson in this dissertation. The first two claims are often attributed to him and form the basis of what some characterize as his "reversal" of the traditional or metaphysical approach to responsibility (see Watson 1987; Chapter 2 for analysis). Strawson never explicitly endorses the last claim, and I am attributing a stronger claim to him than perhaps most would. I am taking it from R. Jay Wallace (1994), whose Strawsonian-inspired analysis of responsibility has been every bit as influential to my interpretation of GM II as Strawson. Wallace's account is especially germane to GM II because there Nietzsche is focused on the morality of fairness or justice, and he argues that punishment, blame, and responsibility bear an essential connection to this aspect of morality. The issue of fairness is also central to Wallace's analysis of responsibility, which he frames in the following terms, "What are the conditions that make it morally fair for us to adopt the stance of holding people responsible?" (Wallace 1994: 5).

By characterizing responsible agency as an issue of competence rather than free will, I do not mean to imply that an analysis of free will is completely *irrelevant* to understanding the nature of conditions of moral responsibility.

For one thing, I understand moral competence to be a kind of *control*, an ability to "grasp and apply moral reasons, and to govern one's behavior by the light of such reasons" (Wallace 1994: 1). One may choose, if they so wish, to define this control in terms of compatibilist free will, though I argue Nietzsche himself did not.[2] The position I attribute to him might therefore be characterized as a kind of "revisionism" about free will and responsibility,[3] rather than a straightforward argument for compatibilism. Secondly, on my reading of GM II, while Nietzsche was very much concerned with the issue of exemption, he shows relatively little concern for the issue of *excuses*, both of which are necessary to Strawson's analysis of responsibility. As noted above, the issue of exemption revolves around whether an agent is an apt recipient of the moral expectations and attitudes that constitute moral blame and responsibility. The issue of excuses, by contrast, concerns the more *particular* issue of whether the agent acted in a way that actually constituted a *transgression* of those expectations, presuming that she is *not* exempt from those expectations and attitudes. On Strawson's analysis, this hinges

---

As we will see, a consequence of framing responsibility in these terms is that the expectation for "goodwill or regard" that Strawson takes to be central to morality plays much less of a role on Nietzsche's analysis. Indeed, I am sympathetic to Leiter's (2019b) argument that Nietzsche would have considered Strawson among the "English psychologists" he derides at various points in GM for this reason.

[2] See Chapter 2, but the main evidence of this is that free will plays no substantive role in Nietzsche's analysis of responsibility in GM II, and indeed is neglected almost entirely. GM II is instead preoccupied with the development of conscience and guilt.

[3] See Vargas (2007).

on whether the agent acted from a morally objectionable quality of will, and he apparently believed that free will *was* necessary to determining this second issue.[4]

I do not attribute to Nietzsche a complete theory of responsibility in this dissertation, because I do not attribute to him a general theory of excuses or excusing conditions. I believe such a theory might be reconstructed from GM II's analysis of responsibility, and I take it Clark (2015) has already gone some of the way in doing so, but I doubt that it was an issue Nietzsche was particularly interested in. Moreover, my investigation of GM II has led me to the conclusion that Nietzsche believes excusing conditions are for the most part cultural artifacts, and eventually come to be influenced a great deal by religious and metaphysical presuppositions we now have good reason to reject (chief among these is libertarian free will).[5] In any case, I take it to be an unsettled question whether free will might have some necessary role to play in a Nietzschean theory of *moral* responsibility, though I am dubious.[6]

As I stated above, the reason for this is that, according to the picture Nietzsche presents in GM II, responsibility is a matter of *competence* rather than free will, because what defines us as responsible agents is our possessing a certain form of *conscience*, not having the "ability to do otherwise" or our possessing compatibilist "source freedom" (McKenna and Pereboom 2015: 39). It is this form of conscience—"bad conscience"—and the agential capacities it confers, which renders blame of oneself and others appropriate. Indeed, if I am right, Nietzsche's view would seem be that the practice of

---

[4] As various of Strawson's exculpatory "pleas" imply: "He couldn't help it," "He was pushed," "He had to do it," "It was the only way," "They left him no alternative" (1962: 77).

[5] See Clark (2015: 90-96). All of this evidence is to be found in Nietzsche's account of the origins of retributive punishment (GM II: 4, 5, 9, and 10), which I summarize in Section 2.2 below.

[6] Note, this should not be confused with the issue of whether free will and autonomy are important to his more general theory of agency and responsibility, especially as they concern the issue of *trust* and the sovereign individual. See Chapters 1 and 5.

blame went awry precisely when we started positing libertarian free will as a *condition* of responsibility. I expand on these remarks in Chapters 1 and 2, which compose Part I of this dissertation. Part I is focused on Nietzsche's views on moral agency and free will, spanning from the first of his critical works on morality, *Human, All too Human* (1878), to two of the writings of his last productive year, *Twilight of the Idols* and *Ecce Homo* (1888). More precisely, Chapter 1 is concerned with Nietzsche's doctrine of "fatalism" and its implications for free will. There I argue, contra Leiter (2015), that Nietzsche's fatalism does not entail that he was an incompatibilist.

In Chapter 2 my attention shifts to GM II, where it remains for the rest of the dissertation. In Chapter 2 I try to address an Interpretive Challenge raised by GM II's naturalistic analysis of responsibility and conscience. Leiter is right that Nietzsche, throughout his corpus, from the first of his critical works on morality to the last, consistently and quite vehemently denies that humans possess (a certain kind of) free will that purportedly makes us responsible. The *Genealogy* is no exception to this general line of criticism, and indeed emblematic of it (GM I: 13). Why, then, does Nietzsche offer a "long history of the origins of *responsibility*" in GM II, and then proceed to praise the "sovereign individual" as a paragon of responsibility? That is, how can we *reconcile* Nietzsche's aims in GM II with his persistent and emphatic denial of free will and moral responsibility throughout his other works? That is the Interpretive Challenge; I argue that the best way to solve it is to interpret Nietzsche as a Strawsonian.

Specifically, I argue that Nietzsche advances his own version of the "reversal" thesis that some have attributed to Strawson, according to which the justification of the moral expectations and attitudes that constitute moral blame and punishment are to be sought *within* those practices, *not* according to independently established a priori

metaphysical conditions. Succinctly put, what it means to *be* responsible is constituted by the practice of *holding* oneself and others responsible. Part II of this dissertation, consisting of Chapters 3-5, provides an elaboration and defense of this thesis. It does so by offering a comprehensive analysis of GM II's account of the development of conscience, guilt, and responsible agency. Below I offer a summary of my findings.

II. "The Long History of the Origins of *Responsibility*"

GM II is indeed a *long* and *complex* "history" of responsibility, one that runs the full gamut of the attitudes and practices that contributed to our becoming responsible agents endowed with a moral conscience. Conscience, Nietzsche tells us, has a "long history and metamorphosis" (GM II: 3). His story begins all the way back in a nebulous state of human "prehistory," when non-moral blame and punishment predominated among humans in ways similar to non-human animals today.[7] Back then, in the "morality of custom" (GM II: 3), we were not yet moral agents, though that primitive form of life did make us *reliable*, or "regular" and "predictable" in our behavior (GM II: 2). It did so because, consistent with Strawson's idea that our social practices are crucially informed by our expectations and attitudes, we held one another to expectations to follow various rules, experienced anger when others violated them, and punished them for doing so. I will refer to this social dynamic throughout and often as a kind of stance, the stance of *holding others to expectations*. This stance is the bedrock of Nietzsche's genealogy of conscience, for without it conscience never would have developed and evolved.

2.1 Expectations of Conformity (Chapter 3)

It is important to distinguish this stance of holding others to expectations from simply making predictions. For instance, I expect my wife to fix her hair each morning, and my

[7] Especially primates, as I argue in Chapter 3.

daughter to wake me up promptly at 6 AM, but in doing so I am just forming a belief about their future behavior. The stance I am concerned with instead involves holding others to *affectively charged* expectations. On Nietzsche's analysis, when we adopt this stance toward other agents, we experience "reactive affects" (GM II: 11), like anger and revenge, whenever these expectations are transgressed. As we saw above, Strawson conceives of the "reactive attitudes" in a similar but narrower way, as responses to the expectation or demand for goodwill or regard. As he understands these, they are essentially *moral* attitudes and expectations, because we hold them toward a person we view as a "term of moral relationships" (Strawson 1962: 86). On Nietzsche's picture, expectations and attitudes of this form would presume the existence of bad conscience and the ability to feel guilt. Nietzsche's "reactive affects" therefore depend on and form a broader class of expectations and attitudes than Strawson's "reactive attitudes." As I will suggest in Section 2.3 below, I believe this to be a perk of Nietzsche's analysis because it helps to buttress Strawson's argument against determinism.

When conscience originated in the morality of custom, it was a *non-moral* faculty. It was not in any way connected to or responsible for the production of moral emotions (e.g., guilt), but it was nevertheless a novel and distinctly *social* "memory," specifically, a "consciousness" or "awareness" of the expectations of others, of the rules they imposed on us. According to Nietzsche, morality originates within this ancient and primitive dynamic in which rules are created and enforced, simply because they are "basic requirements of social co-existence" (GM II: 3). Strawson agrees; he calls this the "minimal interpretation" of morality:

> Now it is a condition of any social organization, any human community, that certain expectations on the part of its members should be pretty regularly fulfilled; that some duties, one might say, should be performed, some obligations acknowledged, some rules observed. We might begin by locating the sphere of

> morality here. It is the sphere of observation of rules, such that the observance of some such set of rules is the condition of the existence of society. This is a minimal interpretation of morality. (1961: 5)

Strawson's "minimal interpretation" of morality and Nietzsche's "morality of custom" are similar in striking ways. Unsurprisingly, it turns out that Nietzsche's interpretation is a lot more violent. In any case, conscience originates within and is uniquely responsive to this social world, one in which some agents have the power to punish and hold other agents to expectations, thus creating a rule or prohibition against that behavior.

I call these *expectations of conformity*. Their violation elicits anger in the punisher toward the transgressor. These expectations are not only ultimately responsible for creating the faculty of conscience, but they also explain the origin of non-moral obligation on Nietzsche's analysis. I argue in Chapter 3 that Nietzsche conceives of non-moral obligations as non-hypothetical imperatives, inescapable "oughts" which lack normative authority. Expectations of conformity represent the first stage in Nietzsche's "long history" of responsibility.

2.2 Expectations of Redress (Chapter 4.1)

Nietzsche's broader developmental approach has various advantages, as I just noted, but it also complicates matters considerably. For on his analysis we can speak of *non-moral* blame and punishment just as we can speak of *moral* blame and punishment—both are "social practices" constituted by their respective expectations and attitudes. In fact, GM II is focused far more on the issue of punishment than it is blame. However, I follow Clark (2015: 90) in taking Nietzschean punishment to involve implicit judgments of blame, and I also follow her in taking him to be concerned not only with explaining the origins of punishment as a natural practice, i.e., one essential to social coordination, but also with the issue of whether an agent is ever *deserving* of punishment.

At a certain point Nietzsche argues that punishment became *retributive* in character (GM II: 4, 10). Retributive punishment is marked by three features on Clark's analysis. It is a backward-looking, non-consequentialist response to wrongdoing; it involves the belief that the offender deserves the punishment he receives, because it is a "just" and "fair" response to a *debt* he owed society; and the offender is allowed to remain a member of society by "paying off" this debt. This practice marks an important transition in Nietzsche's understanding of the concept of obligation and his "long history" of responsibility. Punishment that occurs as a result of violating an expectation of conformity may be and often is arbitrary and wanton, because it is subject to no standards of propriety whatsoever. (See Nietzsche's description of "exile punishment" in GM II: 9.) However, once rules are conceptualized as debts that must be paid off through the principle of "equivalence" (GM II: 4, 8), the idea that offenders deserve to suffer *in proportion* to the harm they caused, punishment becomes subject to internal and culturally variable standards of propriety, and it starts to look more like a moral practice.[8]

Indeed, on Clark's interpretation, Nietzsche's analysis of retributive punishment qualifies as a "primitive form of moral address" (2015: 93). She particularly stresses the fact that agents are now allowed to remain members of the community through punishment, because "retributive ideas … work to isolate the criminal from his deed, so only the deed must be repudiated" (2015: 95). This is significant because it implies that we no longer take the objective stance toward wrongdoers, but rather something that at least looks *like* the participant attitude. I agree, though it is hard to say exactly what this attitude is, for reasons I will expand on momentarily. In any case, I would add that

[8] Importantly, the community now considers whether the criminal's action was "intentional," "negligent," or "accidental" when meting out punishment (GM II: 4), considerations which Nietzsche argues are not present in exile punishment.

retributive punishment is a moral practice on Nietzsche's analysis for at least two additional reasons. First, because punishment is now linked to considerations of fairness and justice, which are moral ideas, and secondly, because once the concepts of fairness and justice modify our social expectations and practices, it is possible for agents to experience resentment.[9]

Therefore, retributive punishment rests on what we might call *expectations of redress*, not expectations of conformity. These expectations of redress are conditioned by the ideas of justice, fairness, equivalence, and desert. Now an obligation is not just a rule one must conform one's behavior to, an inescapable "ought" which lacks normative authority, but as a debt one "owes" it to society to repay, or a requirement that one is legitimately bound to. This marks the second stage in Nietzsche's "long history" of responsibility.

2.3 Moral Expectations, Bad Conscience, and Guilt (Chapter 4.2-6)

Though I endorse Clark's interpretation of retributive punishment in Chapter 4, there I also bring attention to one shortcoming of it. She does not consider how bad conscience and guilt would have modified this practice, and so from the point of view of Nietzsche's "long history" of responsibility, her analysis remains incomplete. (In fairness to her, it was not her intention to offer a complete analysis in Clark [2015].) This is significant on my reading because retributive punishment, as she and Nietzsche describe it (GM II: 4, 14-15), operates solely in accordance with considerations of *objective guilt*, that is, according to whether an offender is deserving of punishment in the eyes of *others*. It does *not* presume the offender would or is even capable of making such a judgment

[9] Resentment is not just a response the expectation or demand for goodwill or regard. It may also be a response to cheaters, free-riders, or anyone who takes herself to be an exception to rules that legitimately apply to her. See Wallace (1994), Walker (2006), Vincent et. al (2018). De Waal (1996) argues that primates have a "sense of social regularity," which is analogous to the sense of justice or fairness, but it does not generalize to other members of one's group.

against *herself*. And that is significant because anyone *incapable* of making this judgment on Nietzsche's analysis is not a moral or responsible *agent*.

Moral agents have a unique comportment to social norms. They *relate* to them in a particular way. Precisely, they view them as "personal" obligations (GM II: 8) that bear on their sense of worth. (Guilt, for Nietzsche, is "the feeling of personal obligation" [GM II: 8]). A person who steals while knowing others will hold her to an expectation of redress, does so in the recognition of her potential indebtedness, i.e., in recognition of the fact that others will view her action as unfair and blame and punish her for it. However, if she had not developed the faculty of bad conscience, the thought "I should not have done that" (GM II: 15) never would have occurred to her. The reason for this is that she does not relate to being fair in the relevant way—being *un*fair doesn't engage her sense of worth. This virtue is not constitutive of her "practical identity," one might say. Why?

Bad conscience is necessary for the judgment "I should not have done that" in two ways. First, bad conscience is created when our aggressive drives can no longer be discharged outwardly and become internalized, or redirected toward the self. This process gives "depth," "breadth," and "height" to the human soul (GM II: 16). This creates the possibility of inner conflict in human beings—the conflict between, say, wanting to be fair and wanting to steal (see Chapter 4, §3.3). Secondly, the judgment "I should not have done that" is an *expression* of bad conscience, a "pang of conscience" (GM II: 14), or guilt. Therefore, an agent who is incapable of undergoing an episode of bad conscience is incapable of making this judgment. This judgment, I argue, is also the expression of one who *holds herself responsible or accountable* (see Chapter 4, §5).

What we arrive at again is a deeply Strawsonian point—perhaps a deeper Strawsonian point than even Strawson himself made (!). Guilt is constitutive of the practice of moral blame, because without guilt moral blame of others would be unintelligible. If a person could not feel guilt, your resentment might appear to her like anger, and it might register as fear, but it would not have its intended effect of inducing an episode of bad conscience, or eliciting guilt. This observation goes some way in defending, I think, Strawson's central argument, namely, that our moral practices are not in need of an "external" or a priori justification, because they are constituted by our expectations and their corresponding attitudes.

Note, the expectations have changed and acquired a new object. When I hold another to these expectations, I do so under the presupposition that she ought to feel *guilt* for violating them. Therefore, we now have the stance of *holding others to moral expectations*. However, my doing so presumes that she can and would *hold herself* responsible. On Nietzsche's analysis, this means she would punish herself by undergoing an episode of bad conscience. Thus, on the view I develop, moral blame and holding others responsible requires that they possess the ability to *hold themselves* responsible.[10]

Nietzsche's account of guilt is much more complex than this, but I hope this brief sketch provides some context to the controversial claim I made above, namely, that being an apt target of the expectations and attitudes constitutive of moral blame is a matter of moral *competence* rather than free will. Bad conscience and guilt mark the third stage in Nietzsche's "long history" of responsibility.

[10] This view is also defended by Paul Russell (2011), who criticizes Strawson for neglecting to provide an account of moral capacity that would show determinism doesn't undermine ascriptions of responsibility.

2.4 The Memory of the Will, Integrity, and Trust (Chapter 5)

Nietzsche opens GM II by asking what it would take to "breed an animal that is permitted to promise" (GM II: 1). I argue that having this "permission" is a metaphor for trust, and that it is underwritten by the last form of conscience we will consider, which Nietzsche calls the "memory of the will" (GM II: 1). On my reading this isn't actually a discrete form of conscience from those we've considered thus far, but rather the maturation and combination of all of them, considered under a different aspect. By calling the conscience a "memory of the will," Nietzsche means to emphasize that it also becomes a *capacity to act*, more precisely, a capacity that the emergence of *bad conscience* made possible. Nietzsche describes this as an ability to "will on and on something one has once willed" (GM II: 1), or as an ability to sustain commitment through a kind of effort of will alone. In short, conceived of as the "will's memory," the conscience is an ability to extend practical commitment in the absence of external incentives. This presumes that the agent instead does so from an internal motive that stems in some way from her values (see 5.2, 5.4).

I argue that the memory of the will underwrites two forms of trust, *moral trust* which has its basis in guilt, and *sovereign trust* which has its basis in integrity. When Nietzsche praises the sovereign individual as a paragon of "responsibility" and "autonomy" (GM II: 2), he does so because he thinks sovereign promisers are rare with respect to the way they constitute and bind themselves by making promises. Indeed, I argue that Nietzsche believes most of us are not "permitted to promise" in the sense that we are not actually *trustworthy*, according to the standards of sovereign trust, and that he thinks anyone who makes a promise purports himself to be so deserving. Consequently, trust which is extended and kept on the basis of moral obligation turns

out to be a kind of weak or degenerate form of promise-keeping on Nietzsche's analysis, though it does not impugn moral trust for all that.

III. Conclusion

In GM II, Nietzsche agrees with Strawson that what it means to *be* responsible is constituted by the practice of *holding* oneself and others responsible, by the expectations and attitudes involved in taking the participant stance toward oneself and others. Nietzsche's "reactive affects" form a broader class of expectations and attitudes than Strawson's "reactive attitudes," but at each stage in Nietzsche's "long history of responsibility" those attitudes and expectations are necessary to explain, and indeed render appropriate, the relevant stance. This occurs in every case because there is a form of *conscience* that has developed and is uniquely responsive to those expectations and attitudes.

1. Non-moral conscience as a memory of "I will nots" is uniquely responsive to holding others to expectations to follow rules.
2. Conscience as a memory of "debts" is uniquely responsive to considerations of fairness and justice.
3. Bad conscience is uniquely responsive to ascriptions of *moral* blame, which is intended to elicit guilt.
4. And the "memory of the will" is uniquely responsive to the need to rely on others from the participant stance, i.e., trust.

Why would an "immoralist" who is so skeptical of free will provide such an account?

Part of the explanation, I think, has to do with the fact that Nietzsche was very much concerned with the issue of moral motivation throughout his career. This is uncontroversial. From the first of his critical works on morality, *Human, All too Human* (1878), where he argued that acting from a moral motive was impossible (see HA I: 34, 133), and thus that we ought to "deny" morality or be immoralists, to *Daybreak* (1881) where he argued that moral motivation was possible but merely a habit inculcated in us

through morality of custom (D 9), and thus permeated with "errors" and so we ought to be immoralists, to the *Genealogy* (1887). There he instead argues that bad conscience and guilt make moral motivation and moral agency possible, but they also make us "sick." Again, this is uncontroversial.

What *is* controversial is whether having the ability to act from moral motives and relate to moral norms in a peculiar way makes us *responsible*. I argue that it does. I argue this because I deny, and I believe Nietzsche denied, that moral agency is the same thing as "free agency." That brings me to the second reason I think an "immoralist" who is skeptical of free will might nonetheless provide a genealogy of conscience and moral agency. Perhaps he did so because, like Strawson, he wanted to *naturalize* our idea of responsibility—not to save it from the "optimists" and "pessimists" who each in their own way "overintellectualize the facts" (Strawson 1962: 91). But because, as he himself claims, he thought we needed "knowledge of the conditions and circumstances" out of which our moral practices grew (GM P: 6), to uncover precisely where they went awry, which in this case requires precisely *not* denying their existence in the first place. This, too, is a fact that Strawson helps us to appreciate.

References

Works by Nietzsche

BGE: *Beyond Good and Evil*. 1886. In W. Kaufmann, trans. and ed., *The Basic Writings of Nietzsche*. New York: Modern Library Edition, 2000.

D: *Daybreak*. 1881. R.J. Hollingdale, trans., Maudemarie Clark and Brian Leiter, ed. New York: Cambridge University Press, 1997.

HA: *Human, All Too Human*. 1878. R.J. Hollingdale, trans. Cambridge: Cambridge University Press, 1996.

GM: *On the Genealogy of Morality*. 1887. Clark, Maudemarie, and Swenson, Alan J, t rans. Indianapolis: Hackett Publishing Company, 1998.

TI: *Twilight of the Idols*. 1888. In W. Kaufmann, trans. and ed., *The Portable Nietzsche*. New York: Viking Press, 1954

Other Works

Clark, Maudemarie. 2015. "Nietzsche on Free Will, Causality, and Responsibility." In *Nietzsche on Ethics and Politics*, 75-96. Oxford: Oxford University Press.

De Waal, Frans. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge: Harvard University Press.

Leiter, Brian. 2019a. *Moral Psychology with Nietzsche*. Cambridge: Cambridge University Press.

__.2019b. "The Innocence of Becoming: Nietzsche against Guilt." *Inquiry* 62 (1): 70-92.

__.2011. "Who is the 'Sovereign Individual'? Nietzsche on Freedom." In Simon May ed., *"Nietzsche's On the Genealogy of Morality": A Critical Guide*. Cambridge: Cambridge University Press.

Russell, Paul. 2011. "Moral Sense and the Foundations of Responsibility." In R. Kane ed., *The Oxford Handbook of Free Will: Second Edition*. Oxford: Oxford University Press, 199-220.

Strawson, P.F. 1962. "Freedom and Resentment." In G. Watson ed., *Free Will: Second Edition*. Oxford: Oxford University Press, pp. 72-93.

__. 1961. "Social Morality and Individual Ideal." *Philosophy* 36: 1-17.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

Watson, Gary. 1987. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In Watson ed., *Agency and Answerability*. New York: Oxford University Press.

Vargas, Manuel. 2007. "Revisionism." In *Four Vies on Free Will*. Malden: Blackwell Publishing.

Vincent, Ring, Rebecca, Sarah, and Andrews, Kristin. "Normative Practices of Other Animals." In A. Zimmerman, K. Jones, and M. Timmons, eds., *The Routledge Handbook of Moral Epistemology*. New York: Routledge, pp. 57-83.

Walker, Margaret Urban. 2006. *Moral Repair: Reconstructing Moral Relations after Wrongdoing*. Cambridge: Cambridge University Press.

Part I

Free Will and Moral Agency

Chapter 1

Nietzsche on Fatalism and Free Will

Abstract: According to Brian Leiter, Nietzsche is a "fatalist" in the sense that he thinks our character and personality are fixed in virtue of natural facts about us (so-called "type facts") and that such facts "significantly circumscribe" the range of possibilities any one person can realize in his or her lifetime (Leiter 2002: 83). What Nietzsche actually endorses, on Leiter's reading, is Causal Essentialism (CE), the view that every substance has a causal essence that constrains its future trajectories, though this essence does not uniquely determine the path that a life must take. In this chapter, I critically assess Leiter's interpretation of Nietzschean fatalism and I argue that it does not have the strong, and, in particular, *incompatibilist* consequences for free will and moral responsibility that Leiter contends. On Leiter's interpretation, Nietzsche's fatalism was developed from, and an extension of, Schopenhauer's fatalistic views concerning character and free will. While Nietzsche did indeed endorse Schopenhauer's view of character and his conception of free will in *Human, All Too Human* (1878), by *Daybreak* (1881) he had rejected both Schopenhauer's unchangeability thesis regarding character, as well as his incompatibilist, leeway conception of freedom. In fact, there Nietzsche began endorsing a compatibilist, source conception of freedom (D 560), in relation to our ability to modify the drives and dispositions that constitute our character, a conception which he endorses well into the works of his last productive year (TI IX: 38). Thus, Nietzsche's fatalistic pronouncements do not show that we lack free will entirely and bear no responsibility for ourselves. In fact, Nietzsche's positive conception of freedom presumes that we can take responsibility for the self we are fated to have.

I. Introduction

Avowals of fatalism are prominent throughout many of Nietzsche's writings, from *Human, All Too Human* (1878), to the works of his last productive year, *Twilight of the Idols* and *Ecce Homo* (1888). These declarations are significant both because fatalism is important to understanding Nietzsche's critique of free will and moral responsibility, as well as his positive ideal of life-affirmation: "amor fati," the "love of fate" (see EH II: 9; GS 276; TI VI: 8, TI IX: 49). My focus in this chapter will be on the critical dimension of fatalism in relation to free will and moral responsibility.[11] The following is one such declaration representative of the connection Nietzsche believes them to share.

[11] For a recent treatment of amor fati, see Stern (2013).

> What alone can *our* doctrine be?—That no one *gives* man his qualities—neither God, nor society, not his parents and ancestors, nor he himself. (The nonsense of the last idea was taught as "intelligible freedom" by Kant—perhaps by Plato already.) No one is responsible for man's being there at all, for his being such-and-such, or for his being in these circumstances or in this environment. The fatality of his essence is not to be disentangled from the fatality of all that has been and will be … One is necessary, one is a piece of fatefulness, one belongs to the whole, one is in the whole. There is nothing which could judge, measure, compare, or sentence our being, for that would mean judging, measuring, comparing, or sentencing the whole. But there is nothing besides the whole. That nobody is held responsible any longer, that the mode of being may not be traced back to a *causa prima* … that alone is the great liberation; with this alone is the innocence of becoming restored. The concept of "God" was until now the greatest objection to existence. We deny God, we deny the responsibility in God: only thereby do we redeem the world. (TI VI: 8)

Above, Nietzsche suggests that we have an "essence" which was "fated" and that our essence is necessarily connected with everything else that has been "fated" in this same way, and that as such both the "whole" and one's place within it are completely "necessary." A consequence of this, or so it would *seem*, is that no one is "responsible" for who she is or what she does, and so no one should be "held responsible any longer." After all, if what we become and what we do is *unavoidable*, then it would seem no one is or ought to be responsible for who one is or what one does.

This is how Brian Leiter invites us to interpret this and other passages where Nietzsche makes fatalistic assertions. The lesson we are to draw from such passages is that we do not possess free will in a sense that would be necessary to justify ascriptions of responsibility, because this would require our being a *causa sui* (self-cause) and possessing free will "in the superlative metaphysical sense" (BGE 21).[12] Leiter moreover argues persuasively that Nietzsche's fatalism was greatly influenced by Schopenhauer, who also espoused fatalistic views of agency and character while endorsing an

[12] Throughout this chapter I will follow Leiter's lead in treating free will as a necessary condition for moral responsibility, though this connection will prove problematic for reasons I expand on in Chapter 2.

incompatibilist conception of free will as the "*liberum arbitrium indifferentiae*," the "free choice of indifference" (1818: 316). What follows here is both a critical analysis of Leiter's understanding of Nietzsche's fatalism and an examination of the trajectory of Schopenhauer's influence on Nietzsche's thinking on these issues.

I argue that Nietzsche's fatalistic pronouncements do not, in fact, have incompatibilist implications for free will and moral responsibility, as Leiter contends. While Nietzsche did endorse Schopenhauer's fatalistic views of character and free will in *Human, All Too Human* (1878), in *Daybreak* (1881) he began deviating considerably from Schopenhauer's views on both points. Whereas Schopenhauer believed our character to be innate, constant, and unalterable (1839: 49-55), Nietzsche flatly rejects that we cannot alter our character in *Daybreak* (D 109, D 560), a reversal from his position in *Human, All Too Human* (HA I: 41). Moreover, he began endorsing a compatibilist conception of source freedom (D 560), in connection to our ability to modify the drives and dispositions that constitute our character, while nonetheless preserving Schopenhauer's psychological determinism. Thus, Nietzsche's fatalistic pronouncements do not show that we lack free will entirely and bear no responsibility for ourselves. In fact, I argue that the most important lesson of Nietzsche's fatalism is that we can and should *take* responsibility for ourselves, and by doing so we realize Nietzschean autonomy.

II. Nietzschean Fatalism

Nietzsche's fatalistic pronouncements are rather idiosyncratic and do not easily lend themselves to a worked out philosophical position. A fatalist, Richard Taylor tells us, believes he cannot do anything about the future, or that its occurrences are not within his power to do or to forego, and so the fatalist has the same attitude toward his future as we

all have toward our past (1962: 56). Understood in these terms, fatalism may denote a *belief* about the future and/or a kind of *practical attitude* in response to this belief. In its general form, fatalism is the belief is that all events are, and always have been, *non-contingently unavoidable*.[13] More strictly, the fatalist believes that for every person P, action A, and time T, P is able to A at T (or P has the power to A at T) only if P actually A's at T. It follows from this definition that everything we do (or fail to do) we do (or fail to do) with necessity—we are not able to do otherwise.[14]

I will now consider two different versions of fatalism that have been attributed to Nietzsche. The first version, offered by Lanier Anderson (2013), I consider to be closer to Nietzsche's considered view, despite the fact that it does not generate a particularly threatning notion of necessity, because it serves Nietzsche's broader practical aims. The second version I will consider is offered by Brian Leiter (2002), and it will be my focus throughout the remainder of the chapter.

2.1 Fatalism as Inverse Super Essentialism (ISE)

The passage above from *Twilight of the Idols* seems to be a good representative of the strict definition of fatalism. So, too, is GM I: 13, where Nietzsche claims:

> To demand of strength that it *not* express itself as strength…is just as nonsensical as to demand of weakness that it express itself as strength. A quantum of power is just such a quantum of drive, will, effect—more precisely it is nothing other than just this driving, willing, effecting itself … small wonder if the suppressed, hiddenly glowing affects of revenge and hate exploit this belief and basically even uphold no other belief more ardently than this one, that *the strong one is free* to

[13] Thank you to Michael Nelson and Eric Schwitzgebel for pressing me to clarify fatalism, to distinguish it from causal determinism. As Eric suggested to me, the difference can be captured as follows. Determinism implies that *if* Oedipus had not run away from home, he would not have married his mother. Whereas fatalism implies that, *even if he had not* run away from, he still would have married his mother. According to fatalism, Oedipus' action was non-contingently unavoidable, whereas according to determinism it was only contingently necessary.

[14] This definition was suggested to me by Michael Nelson (personal corr.).

> be weak, and the bird of prey to be a lamb:—they thereby gain for themselves the right to hold the bird of prey *accountable* for being a bird of prey … (GM I: 13)

The idea that Nietzsche clearly means to convey here is that the bird of prey is not "free to do otherwise." This has something to do with the fact that, being strong, the bird of prey *must* express itself "as strength." It *cannot* express itself "as weakness," as the lamb would like. Why?

According to Lanier Anderson, in the above passage Nietzsche is endorsing *inverse super essentialism* (ISE), the view that

> [T]here is *no* complete concept or essence of a thing separate from its properties and effects, so the thing itself is nothing but their collection. Each property is thus necessary in the degenerate sense that without it, the *collection* of properties would be different (2013: 436).[15]

Applied to persons, ISE holds that "*every single one* of a person's actually existing (i.e., past and present) properties (whether they are actions or not) is necessary" (Anderson 2013: 435). ISE takes events or properties to be "necessary" based on the following principle of identity: "the mere fact that things are what they are and not anything else" (Anderson 2013: 435-6).

So, according to ISE, things are *necessarily* what they are, because if they were any different they would not *be* what they are. That is why the bird of prey *must* express itself "as strength." However, Anderson believes that ISE's derivative notion of necessity is rather tame. As he explains:

> [T]he *sort* of necessity [Nietzsche] attributes to events is remarkably weak. For him, "event and necessary event is a *tautology*" (WP 639). If the concepts <event> and <necessary event> are really just *identical* (tautologous), then it follows that all events are necessary—but also that necessary events are just

[15] Anderson also adduces as evidence in favor of this view a number of passages from Nietzsche's unpublished notes (see WP 551, 552, 557, 634, 639).

> events (construed as lacking any core essential features that could preserve their identity across counterfactual change). That is, things are necessary *only* in the degenerate sense that they are what they are and not something else: apparent necessity is "only an *expression* for the fact that a force is not also something else" (WP 552). (Anderson 2013: 436-7)

So, does ISE imply fatalism? Does it imply that P is able to A at T (or P has the power to A at T) only if P actually A's at T? Yes, but only in the sense that if P did *not* A at T, P would not *be* P. Because of this, Anderson thinks that ISE only implies fatalism about the past and present: "But the future *is* not anything (yet). Therefore, Nietzsche has no grounds for insisting that it *necessarily be* one way or another" (Anderson 2013: 437).

Why? Because, according to Anderson, Nietzsche rejects causal necessity in a number of passages, the thesis of nomic determinism explicitly, and so "there is nothing left to determine the future, and it remains *open*" (Anderson 2013: 436). According to nomic determinism, an event E at time T is determined by conditions C of the remote past in conjunction with the laws of nature L. So, E, including some choice or action made by an agent, is *contingently necessary*, per nomic determinism, assuming both C and L. If C and L obtain, the agent cannot do otherwise than she in fact does. Nietzsche's skepticism about nomic determinism is directed primarily at the idea of their being "natural laws," particularly the idea that such laws ensure that causes *necessitate* their effects. In one context he suggests that this inference relies on a faulty, mechanistic picture of the universe, one that involves a "misuse of 'cause' and 'effect'" (BGE 21). In another he suggests that explanations in terms of natural law presuppose a *lawgiver*, someone with a will (God), who would have ordered the universe in a law-like way: "Let us beware of saying that there are laws in nature. There are only necessities: there is nobody who commands, nobody who obeys, nobody who trespasses (GS 109). In any

case, Nietzsche's specific reasons for rejecting nomic determinism need not preoccupy us here—it's enough for us to take note of the fact that he does reject it.[16]

However, even if this is true, it's not clear why the future should remain open, as Anderson contends. Indeed, this reasoning seems to confuse *causal* necessity with *logical* necessity, or at least to make the latter depend on the former. As he explained, though, ISE is a view about logical necessity; the necessity comes from the principle of identity he articulated—things are *necessarily* what they are only because if they were any different they would not *be* what they are—not from C and L. So, the only way ISE would not entail fatalism about the future is if future properties are (for some reason) *not* included among the set of P's properties. But do we have reason to think Nietzsche would have believed that to be the case? I see no reason to think he would have, his skepticism of nomic determinism notwithstanding, because drawing this inference would involve a seemingly arbitrary restriction on his principle of identity.[17]

Still, Anderson is right to insist that the kind of necessity we get with ISE is rather cheap. (Things are *necessarily* what they are only because if they were any different they would not *be* what they are.) So, we might ask: why would Nietzsche endorse this view of identity? Two reasons come readily to mind. First, ISE is incompatible with a type of freedom Nietzsche unequivocally means to reject: the ability to do otherwise (see GM I: 13). And so, by claiming ISE is true, Nietzsche is in effect telling us that we do not possess libertarian free will. Ridding the world of this erroneous belief in free will is moreover necessary to "restore innocence" to the world of "becoming" (TI VI: 8), because the priests of Christian tradition promulgated a perverse and cosmic notion of

---

[16] For analysis, see Leiter (2002: 22-3) and Clark and Dudrick (2012: 94-7). Leiter argues that, in BGE 21, Nietzsche's skepticism about determinism was due to his Neo-Kantian phenomenalism that he eventually rejected. Clark and Dudrick argue, to the contrary, that BGE 21 expresses a Neo-Humean regularity view of natural laws.

[17] Thanks to Michael Nelson and Eric Schwitzgebel for pressing me to see this.

guilt in connection to this idea of free will. Specifically, they "infected" the natural world with the idea of guilt by interpreting innocent misfortune and bad luck as deserved punishment for one's freely chosen sins, thereby creating a "moral world order" (TI VI: 7, cf. A 24-5; for analysis see Chapter 4, §5).[18] Nietzsche does indeed want us to overcome *this* idea of guilt, as well as the ideas of free will it depends upon.

Secondly, in many of the passages where Nietzsche endorses ISE, he does so in an effort to engender a certain *practical attitude* in himself and his readers. Consider the following passage:

> The fatality of [one's] essence is not to be disentangled from the fatality of all that has been and will be … One is necessary, one is a piece of fatefulness, one belongs to the whole, one is in the whole. There is nothing which could judge, measure, compare, or sentence our being, for that would mean judging, measuring, comparing, or sentencing the whole. But there is nothing besides the whole. (TI VI: 8; see TI II: 2)

Nietzsche's remarks here rests on considerations surrounding identity, not causal necessity. His contention is that we cannot judge an individual life because it cannot be separated from the whole, and we should not judge the whole. Why does he care to make this point? Because he wants to embody amor fati, the "love of one's fate."

> [A] spirit who has *become free* stands amid the cosmos with a joyous and trusting fatalism, in the *faith* that only the particular is loathsome, and that all is redeemed and affirmed in the whole—*he does not negate any more* (TI IX: 49).
>
> I want to learn more and more how to see what is necessary in things as what is beautiful in them—thus I will be one of those who makes things beautiful. *Amor fati*: let that be my love from now on! I do not want to wage war against ugliness. I do not want to accuse; I do not even want to accuse the accusers. Let *looking away* be my only negation! And, all in all and on the whole: some day I want only to be a Yes-sayer! (GS 276).

---

[18] This also provides necessary context to the notion of "responsibility" Nietzsche rejects in this passage. "We deny God, we deny the responsibility in God: only thereby do we redeem the world" (TI VI: 8). The notion of responsibility being rejected in this passage relies on a particular, historically conditioned conception of free will as an undermined capacity for choice. He is not rejecting *all* forms of guilt and responsibility.

To "love" one's fate, a person cannot have any regrets about his life, to wish it had been any different than it in fact turned out to be. To do that, one must come to see *all* the events that have shaped one's life as being *necessary* and *essential* to it. I believe this is because Nietzsche thinks, plausibly, that regretting an event is typically accompanied by the thought that what occurred was merely accidental or contingent—"if only" he had left five minutes earlier, he would not have been in the car accident—which indulges the fantasy that what occurred was not, in fact, *essential* to one's life, and so one's life might be affirmed in its absence. Amor fati instead demands that we affirm our life in its *totality*. And to do that we must see everything within it as being *necessary*. That is, one must affirm ISE.

So, Nietzsche endorses ISE despite its weak notion of necessity, I suggest, because it works in the service of these other practical aims. Fatalism is not the result of decrees of fate made by agents with mystical powers, divine foreknowledge, or causal necessity. Yet, like these, ISE does rule out our having the ability to do otherwise. And, by espousing this belief in fatalism, Nietzsche did *not* mean to engender an attitude of resignation because one is incapable of doing anything to change the future.[19] Like the Stoics (see Rutherford 2011), he wanted to engender an affirmative attitude towards our fated lives, which he calls "amor fati," love of one's fate. For these reasons, I take ISE to be a good representative of Nietzsche's considered view concerning fatalism. However, I

[19] What Nietzsche calls "Mohammedan fatalism": "they think that man will stand before the future feeble, resigned and with hands clasped because he is incapable of effecting any change in it: or that he will give free rein to his impulses and caprices because these too cannot make any worse what has already been determined" (WS 61).

He speaks of Russian fatalism in similar terms: "that fatalism without revolt which is exemplified by a Russian soldier who, finding a campaign too strenuous, finally lies down in the snow. No longer to accept anything at all, no longer to take anything, no longer to absorb anything—to cease reacting altogether" (EH I: 6).

now turn to a competing interpretation of Nietzschean fatalism offered by Brian Leiter, which will be my focus in the remainder of this chapter.

2.2 Fatalism as Causal Essentialism (CE)

Leiter, too, takes it to be a constraint of Nietzsche's fatalism that it is incompatible with, and does not depend on, the truth of nomic determinism. He proposes instead that we understand Nietzschean fatalism in terms of Schopenhauer's fatalistic views of character and free will.

> Following (though modifying) Schopenhauer, Nietzsche holds that a person's life proceeds along a fixed trajectory, fixed by 'natural' facts about that person. Nietzsche, the fatalist, views a person like a plant: just as, say, the essential natural facts about a tomato plant determine its development (e.g., that it will grow tomatoes and not, say, corn), so, too, the essential natural facts about a person determine its development as well. Of course, the precise development of a tomato plant – whether it flourishes or wilts – is affected (causally) by a host of other factors that do not constitute the "essence" of the plant: for example, the soil in which it is planted, the amount of water it receives, and the like. (2002: 81)

Nietzsche, on Leiter's view, is a *causal essentialist*, meaning that he believes every substance has a causal essence that constrains its future trajectories, though this essence does not uniquely determine the path that one must take. Applied to persons, causal essentialism implies that every person has an essence that "significantly circumscribes" her possible life trajectories, and thus that we are all "fated" because what we can become and do "is severely constrained from the start" (Leiter 2002: 83).

As we can see, causal essentialism (CE) is not a thesis about the future being non-contingently unavoidable (fatalism proper), nor is it a view about conditional necessity (nomic determinism), but rather a view about *circumscription*. CE is a view about what I *cannot* do or become, per some antecedent conditions obtaining, not about what *must* occur. As Leiter claims, CE "does not entail that any particular outcome to a person's life

is *necessary*" (2002: 83); it "entails only that one's possibilities are circumscribed" (2002: 98). More formally, CE is the view that:

> For any substance S, S has certain essential properties EP that are *causally primary* with respect to its future.

To be causally primary means that EP are *necessary*, though they may not be *sufficient*, to bring about any given effect (Leiter 2002: 81-2).[20] So, EP "non-trivially determine the space of possible trajectories" for S (Leiter 2002: 83). That is, EP play a circumscriptive or restrictive role: they establish the boundaries of S, of what S can and cannot become and do. For instance, you might think that the fact that I am only 5'8" tall and possess moderate athletic ability ruled out the possibility that I would ever become an NBA basketball player. However, Leiter intends for CE to be a stronger thesis than this. CE is supposed to entail that a person's life is "significantly circumscribed" or "severely constrained" (2002: 82, 83). In order to understand why this is so, we must connect CE to another view Leiter attributes to Nietzsche, which he calls the "Doctrine of Types."

According to the Doctrine of Types, "each person has a fixed psycho-physical constitution, which defines him as a particular *type* of person" (2002: 8). Type-facts are understood by Leiter to be "largely immutable" physiological or psychological facts about persons (2002: 91), and they perform two important roles in his understanding of Nietzsche's moral psychology. First, they are among the principle explanatory tokens Nietzsche appeals to when attempting to explain a person's beliefs, values, and actions

[20] So, for instance, EP that could be causally primary with respect to being an elite marathon runner are things like: having a high VO_2 max, a high lactate threshold, and a slender body. For example, having a high VO_2 max, the ability to process lots of oxygen during strenuous exercise, is *necessary* to be an elite marathon runner, though this alone does not ensure that one *will* become one. Among other things, an elite marathon runner must also run upwards of 80 miles per week, and so she must live in a part of the world that allows for this kind of training, as well as a lifestyle that affords the free time to do it.

(Leiter 2002: 8, 95). For instance, Nietzsche believes human behavior and our consciously held beliefs are best explained in terms of our subconscious drives, but a person's drives on Leiter's views are just "type-facts" about the person. In this capacity type-facts are synonymous with any non-conscious cause, or any cause that is not the effect of a conscious act of will. Secondly, and more narrowly, per the Doctrine of Types a certain *set* of type-facts is supposed to exist for each person such that these facts *demarcate* the individual and categorize him into rigid classes or kinds (e.g., "weak," "strong," "sick," "decadent," "healthy"). In order to play this demarcating role, type-facts must be more than just non-conscious causes; they must be resistant to change, or "fixed," as Leiter says. I will refer to this subset as *rigid type-facts*.

Leiter proposes—again, following Schopenhauer—that "the fundamental facts about one's *character and personality* are fixed by [type-facts]" (2002: 82, see 2002: 58).[21] In other words, he proposes that a person's character and personality are comprised of rigid type-facts. Character and personality traits are invoked to explain why people behave the way they do; they are *dispositions* to behave in various ways (e.g., courageously, shyly, humbly). Such dispositions are moreover taken to be stable, so that the person will act or think in the characteristic ways across differing contexts. This allows for the sorts of attributions that constitute character, which is a socially constructed identity consisting of these dispositions we identify in ourselves our ascribe

[21] Leiter says "natural facts" instead of "type-facts" here, but this leaves open the issue of *which* of those natural facts about our character and personality are essential and causally primary. Leiter later confirms that he has in mind type-facts and not natural facts of another kind (2002: 97-8).

to others. Finally, these character attributions are often *evaluative* in tone, as in the case of virtues.[22]

So, because a person's character consists of stable and enduring dispositions that influence her behavior across time, in different contexts, and throughout her life, these dispositions, if they are "fixed," would indeed severely constrain what a person could become and what she could do. According to Leiter, "how one responds to differing circumstances and environments [would] also [be] causally determined by [these type-facts]" (2002: 82). Note, however, a trait T could be "fixed" in the sense that:

(i) T is innate (part one's essential nature) and immutable; or

(ii) T is innate and recalcitrant to revision, though alterable.[23]

If a trait of character is "fixed" in the sense of (i), it would be like a mountain or a large boulder blocking one's path, something that one can only avoid by going around or perhaps over. For Schopenhauer, traits of character are "fixed" in this sense. If T is "fixed" in the sense of (ii), however, it would be more like a wood plank securely fastened by nails—alterable, though doing so may take considerable effort and may be impossible without the necessary tools. As we will see, Nietzsche eventually comes to hold a view

[22] According to Alfano (forthcoming), Nietzsche's view is that we acquire character either by labelling ourselves or by being labelled by others. For instance, Nietzsche argues that virtues were born when the nobles began labeling themselves "good," according to their dominant character traits, such as their courage or truthfulness, and they labelled those they held in contempt, the plebeians, "bad" according to their dominant character traits, such as their cunning or obsequiousness (GM I: 2, 5).

[23] In the first edition of *Nietzsche on Morality*, Leiter does not tell us what it means for a trait to be "fixed." In the second edition of the book, he does. "Type-facts are 'fixed,'" he says, "in the sense that persons cannot choose the type-facts they have; they are not necessarily 'fixed' in terms of their relative strength or importance at different points in a person's life" (2015: 7). According to this definition, T is "fixed" in the sense that (iii) T is innate, enduring, and not the product of choice. This definition, unfortunately, is ambiguous between (i) and (ii). To say that type-facts cannot be chosen is only to say that they are not the products of choice, and so cannot be things we acquire or change merely by deciding to do so, not that we lack the power to change them.

like (ii), though early in his career he did follow Schopenhauer in believing traits of character were "fixed" in the sense of (i).[24]

CE is the core of Nietzsche's fatalism on Leiter's interpretation, though not the whole story. As noted above, the "essential natural facts" (2002: 81) about a tomato plant ensure that it will grow tomatoes *only* if it receives enough water, sunlight, and nourishment from the soil. So, CE allows for environmental and circumstantial factors that are not themselves the effects of type-facts to play a causal role in determining the future trajectory of some substance as well. Combining these two factors, we arrive at Nietzsche's fatalism:

> At any given moment, the trajectory of a person's life is determined by type-facts plus environmental or circumstantial factors.

Leiter (2007: 7, fn.11) endorses the above definition, originally offered by Owen and Ridley (2003).[25] However, note that environmental and circumstantial factors are causes that are *contingent* and hence introduce luck and *indeterminacy* into Nietzsche's picture of the future. That I was born and raised in Midwest United States, for instance,

---

[24] Leiter contends, to the contrary, that Nietzsche maintains a basically Schopenhauerian, fatalistic view of character throughout his career (2002: 58-63). Much of the evidence he presents in favor of this is equivocal though, because it is also evidence of ISE and/or amor fati, including "the most striking evidence" Leiter offers when discussing fatalistic themes in the second book of *Ecce Homo* (2002: 83). However, a crucial piece of evidence that does seem to favor Leiter's view of rigid type-facts in the sense of (i) is the following passage from BGE:

> Learning changes us … But at the bottom of us, really 'deep down,' there is, of course, something unteachable, some granite of spiritual *fatum*, of predetermined decision and answer to predetermined selected questions. Whenever a cardinal problem is at stake, there speaks an unchangeable "this is I" (BGE 231).

However, Leiter does not provide commentary of the surrounding context and doing so misleads the reader about Nietzsche's intent in this passage. What Nietzsche thinks is "unchangeable" are certain *prejudices* that we harbor, "about man and woman, for example." Directly after this passage, Nietzsche offers a number of highly offensive remarks about "woman as such," acknowledging that they are only "*my* truths" (BGE 231). (See Clark [1994] for analysis.) Thus, one cannot infer from this passage that he thinks traits of character are "fixed" in the sense of (i), though one can conclude from it that Nietzsche believes we all harbor such "truths."

[25] Owen and Ridley mistakenly take this to be a definition of CE when it is instead an *implication* of CE. Curiously, Leiter does not correct their mistake.

has had an indelible impact on the development of my character (e.g., my esteem for modesty and humility), but this is a contingent fact about me, since it is possible that my parents could have chosen to live or relocated somewhere else. So, in what sense is it "necessary" that I value modesty and humility?

As the above definition makes clear, Leiter does intend for Nietzsche's fatalism about the future to entail a kind of necessity, only one that does not depend on the existence of laws of nature. Instead, he proposes that Nietzsche endorses a form of *psychological determinism*, according to which human behavior is caused by our subconscious drives and we are powerless to change the fundamental drives that constitute our character. So, although Nietzsche's fatalism about the future allows for contingencies, since environmental and circumstantial factors are causes which I lack control over, their effects are "necessary." When put in these terms, Nietzschean fatalism can be understood in terms of the following, modified version of the consequence argument (Van Inwagen 1983):

1. No one has power to shape the type-facts and environmental or circumstantial factors that determine one's future.
2. Type-facts and environmental or circumstantial factors, since these exhaust all of the causally relevant factors, entail every fact about the future.[26]
3. Therefore, no one has power over the facts of the future.

[26] Pamela Hieronymi expressed confusion about this premise, which I share. "Entail" here seems to be the wrong word, since environmental and circumstantial factors are not determined according to natural laws. Leiter nonetheless believes these three factors determine everything about the future (Leiter 2007: 7, fn.11). To capture this, I propose we think of "entails" in the following way. These three factors "entail" the future in the sense that the agent is powerless to do anything to change it. Strictly speaking, nothing is determining environmental and circumstantial factors to make it so that they turn out the way they do—it's just a matter of luck or indeterminacy—but from the *agent's* perspective it's all the same.

It bears mentioning that Leiter nowhere endorses the above argument, though I do think he is committed to it. He certainly thinks Premise 2 is true (see Leiter 2007: 7, fn.11). The uncertainty surrounds Premise 1, which will be my focus from here on.

As we will see momentarily, Schopenhauer uncontroversially would accept the above argument. (Actually, he would accept a version of the argument that appealed to laws of nature as well.) He held that a person's character is fixed from birth, constant, and unchangeable. According to him, the best we can do is change the *circumstances* we find ourselves in, so as to avoid those occasions when disagreeable traits will reveal themselves. Thus, we lack entirely for Schopenhauer the freedom or ability to change the constitution of our character. Leiter argues that this sort of external modification is the best we can accomplish on Nietzsche's view as well (see 2002: 62-3, 97-8).[27] However, he is not always consistent about this, since he takes external modification also to encompass modification of our inner drives and dispositions (Leiter 2002: 62-3). So, it is unclear whether he thinks Premise 1 is true. Nevertheless, I think there are three reasons why he should accept it.

The first and most weighty consideration is that he believes Nietzsche is a hard determinist, that is, someone who is both an incompatibilist about free will and affirms the truth of determinism. So, if Leiter rejected the above argument, it is unclear *why* he would and what alternative he would replace it with. The second consideration is that he is committed to their being a strong analogy between Schopenhauer's and Nietzsche's views on agency and character. So if Schopenhuer would accept this argument, on

[27] For instance: "But what we can contribute, qua gardeners, is to shape the *environment* in ways that will affect which of the possible trajectories – wilting, flourishing, or any of the possible stages in between – the plant will realize" (Leiter 2002: 98, emphasis mine).

Leiter's reading Nietzsche ought to as well.[28] The final consideration is Leiter's Doctrine of Types, which would require that character traits are "fixed" in the sense of (i), or else "types" would merely be loose classificatory schemes. Leiter instead intends for them to correspond to "largely immutable" (2002: 91) traits that are effective in producing an agent's actions and beliefs throughout her life (i.e., causally primary).

In what follows, I will argue that Premise 1 is not true for Nietzsche, at least not beginning within *Daybreak* (1881). Starting there, he does not take character to be comprised of traits that are "fixed" in the sense of (i). This raises doubt whether Nietzsche was a proponent of the Doctrine of Types, but more significant for my purposes is the fact that this shows Nietzsche grants humans a *kind* of freedom that Schopenhauer and Leiter deny we have.

III. Schopenhauer's Fatalism

To understand Schopenhauer's fatalism, we must first distinguish between two conceptions of freedom. According to McKenna and Pereboom, proponents of *leeway freedom* believe an agent's will is free if she has a choice between at least two alternatives, the ability to do otherwise (2015: 38). Proponents of *source freedom*, on the other hand, believe an agent acts freely whenever she is the initiating source of her action, whenever she determines her will (perhaps, say, by making a reflective decision or by identifying with some motive that is causally effective) (McKenna and Pereboom 2015: 39). One can be an "incompatibilist" or "compatibilist" about either conception of freedom, depending on

[28] In addition to Leiter's tomato plant analogy (2002: 81), quoted above: "If Nietzsche resists Schopenhauer's pessimism and his moral philosophy, Nietzsche follows him much more closely in his theory of agency and character" (2002: 58).

i) whether one thinks free will requires the ability to do otherwise, or whether it only requires being the source of one's action, or both, and
ii) whether determinism robs us of just one kind of freedom, or both kinds.

I mention the distinction now because it will prove essential to understanding the kind of freedom Nietzsche affirms, and which Schopenhauer rejects. As we will see, both are incompatibilists about *leeway* freedom, but Schopenhauer is also an incompatibilist about *source* freedom, whereas Nietzsche is not.

Schopenhauer, foreshadowing Harry Frankfurt (1969, 1971) many years later, was primarily concerned with internal freedom *of* the will, as opposed to the ability to manifest *what* one wills externally in the world. Early in *On the Freedom of the Will*, Schopenhauer disparages compatibilists, like Hume, who focused narrowly on the "freedom of action," which he calls "*physical freedom*" (1839: 3), arguing instead that the more fundamental problem surrounding free will concerns the "freedom of *willing*," understood to be "the relationship of willing itself to [a] motive" (1839: 16). The fundamental problem of free will, then, is taken by him to be a problem *internal* to the agent and concerns the ways in which the will is moved by other motives or causes, not the way in which action may be constrained by one's external circumstances.

This internal or "*moral freedom*" (1839: 5), as he calls it, is intuitive but hard to define. People generally think they possess moral freedom simply by making choices—"I am free when I can *do what I will*" (1839: 6)—but, as Frankfurt's addicts made plain and as Schopenhauer himself argues, this falls well short of a *proof* of free will, for then we could just ask: "Can you also will that which you will to will?" (Schopenhauer 1839: 6). Schopenhauer, unlike Frankfurt, contends that this question leads to an infinite regress, unless we define free will in "negative" and "absolute" terms as a will that "would not be

determined by anything at all" (1839: 8).[29] Later he calls this the "*liberum arbitrium indifferentiae*," the "free choice of indifference" (1818: 316), claiming it to be "the only clearly defined, firm, and positive concept of that which is called freedom of the will" (1839: 9).

So, the conception of freedom that Schopenhauer thinks is most fundamental is *source freedom*. However, he thinks in order to possess this "moral freedom," the will must be *absolutely* free—the liberty of indifference—to rebut the problem of infinite regress. That is, he thinks we must also possess absolute *leeway freedom*. Free will is completely illusory, on his view, because we do not possess absolute leeway freedom. To see why, we will now consider his theory of action and his views on character.

3.1 Schopenhauer's Theory of Action

Let's begin with what we *are* free to do, according to Schopenhauer. He famously compares our "freedom" to act to the different ways in which water can behave depending on the external cause acting on it.

> "It is six o'clock; the day's work is over. I can now go for a walk, or go to the club; I can also climb the tower to see the sun set; I can also go to the theater; I can also visit this or that friend; in fact I can also run out by the city gate into the wide world and never come back. All that is entirely up to me; I have complete freedom; however, I do none of them, but just as voluntarily go home to my wife." This is just as if water were to say: "I can form high waves (as in a storm at sea); I can rush down a hill (as in the bed of a torrent); I can dash down foaming and splashing (as in the waterfall); I can rise freely as a jet into the air (as in a fountain); finally, I can even boil away and disappear (as at 212 degrees Fahrenheit); however, I do none of these things now, but voluntarily remain calm and clear in the mirroring pond." Just as water can do all those things only when the determining causes enter for one or the other, so is the condition just the same for the man with respect to what he imagines he can do. Until the causes enter, it is impossible for him to do anything; but then he *must* do it, just as water

[29] Interestingly, Schopenhauer here also anticipates Watson's (1975) famous critique of Frankfurt.

> must act as soon as it is placed in the respective circumstances. (Schopenhauer 1839: 43, cf. HA I: 106).

Action is constrained in two ways above. First, the domain of possibilities is circumscribed by the substance's nature or character. Secondly, once the "causes enter" and interact with one's nature or character, the effect follows necessarily (e.g., once the motive to go home to his wife presents itself in consciousness, the man *must* act as he does). According to Schopenhauer, every action a person performs is the necessary consequence of these two factors, of the operating motive in conjunction with one's character (1839: 58).

Schopenhauer conceives of character as the "basis and ground" (2005: 58) upon which motives operate, as a kind of filter, without which they would produce no effect.[30] Consider, first, the case of water. Water cannot burn like paper, boil at 80 degrees Fahrenheit, or flow uphill because it is not in water's *nature* to do so. Just the same, water cannot form high waves, rush down a hill, or boil away unless it was within water's nature to do these things. Similarly, imagine that three people stumble upon a snake in the woods. One person is terrified, another threatened, and the other curious. We can imagine further that the terrified individual cannot bear to look at the snake, the threatened person keeps constant guard over it and is ready to swing a stick at it, and the curious person is attempting to pick it up. What explains these different reactions to the

[30] Schopenhauer conceives of causation as the interaction between "two factors" or forces, one inherent to the thing being acted upon, and one external to it that induces change. As he says, "No cause in the world ever brings about its effect all by itself, or produces it out of nothing … This change always corresponds to the nature of that which, therefore, must already have the force to produce it. Thus every effect originates from two factors, an inner and outer one, namely, from the original force of that which is being acted upon, and from the determining cause which forces the former to manifest itself in this case" (1839: 47-8). Character is the "inner/original force" or "factor" and the motive the "outer force" and "determining cause."

same external stimuli, according to Schopenhauer, is that our character plays a crucial role in determining the effect.

A motive, for Schopenhauer, is a "causality which passes through cognition" (1839: 32). According to Christopher Janaway, Schopenhauer conceives of a motive as a "conscious perception or thought that occurs in a subject's consciousness and causes, or is at least apt to cause, a willed action of that subject" (2012: 439). So, in the case we have just imagined, the terrified person has a motive to flee and avoid the snake, which in conjunction with facts about the agent's character in the circumstances, necessitates that she acts as she does. This, in a nutshell, is Schopenhauer's picture of action: an action is determined by a motive in conjunction with one's character in the circumstances. More formally: *A is determined by M in conjunction with C in X*. Specifically, M is occasioned by X, some outward fact about the world (e.g., encountering a snake) or some fact about the agent (e.g., being hungry), which then gets represented in the agent's consciousness and, based on C, determines her action.[31]

It also bears noting that Schopenhauer, like Hume, believed reason to be subordinate to one's motives. Schopenhauer does grant human beings a "relative freedom," a freedom from "the immediate compulsion of the perceptually present objects which act as motives" (1839: 36), but this amounts to little more than acknowledgement of the fact that we do not act *instinctually*, like other animals. "The necessity of the effect of motives is not in the least obliterated, or even diminished," Schopenhauer contends, since the only thing that reason can do is bring about *indecision* by making

[31] So, for example, the terrified person would not pick up the snake because it is not in her nature to do so. Similarly, someone who does not like a particular food would not eat it, even if it were available. This does not rule out the possibility that, if the circumstances were different (e.g., the person were starving), she would.

salient an underlying "conflict of motives" (1839: 36). He expands on this idea in the *World as Will and Representation*; there he suggests that the intellect's only contribution to making a decision is a "clear display of the motives on both sides" (1818: 318), but reason resides in a "subordinate position" and is a mere "spectator" and "passive" (1818: 317) with respect to any given motive's influence over the will.[32] In other words, the intellect or reason merely serves as the will's "eyes," as it were. Schopenhauer later and illuminatingly speaks of reason's relationship to the will in terms of a "strong blind man carrying a sighted lame man on his shoulders" (1844: 209). Reason, then, does not bring us anywhere close to freedom of the will in the "absolute" and "negative" sense.

3.2 Schopenhauer on Character

As we saw above, we do have the "ability to do otherwise" on Schopenhauer's view, though only in a very restricted and deflationary sense. Consider, again, the case of water. If water can form high waves, rush down a hill, or sit motionless in a pond, then it is within water's nature or character to do all these different things, and so by the analogy water can "act otherwise," as confined by its set of inherent and unalterable dispositions. Of course, this is not a kind of freedom worth wanting, because water has

[32] The will, for Schopenhauer, is "primordial" (1818: 319) metaphysically basic. According to Janaway, Schopenhauer conceives of willing very broadly as "end-directed-activity." "It matters not whether the activity is rational or non-rational, learned or instinctive, conscious or unconscious, or even whether it is the activity of the organism as a whole or merely a sub-function of the organism" (Janaway 2012: 440). Willing in this sense applies to activities like my filling a cup of water in order to drink it, a tree growing taller and producing fruit, and the heart pumping blood. Willed actions differ from these activities only because they are end-directed-activities mediated through the medium of cognition, or consciousness. Specifically, acts of will are "caused by occurrent motives" and so are "the effect of cognitive states which give them their particular aim or content" (Janaway 2012: 440). However, acts of will and willing in general share a common *essence*—that of *striving*. According to Schopenhauer's metaphysics, "Everything in nature is constantly striving, and the common essence of the world is will, in this extended sense of the term" (Janaway 2012: 440). As we will see, striving is also important to Nietzsche's conception of willing.

no *control* when it comes to acting in these different ways; it is only "free" to do so once acted upon from the outside by some cause (e.g., the wind). So, water is itself utterly *passive* when it comes to "acting" in these different ways, because water is unable to change its chemical constitution, and so it is unable to change its nature or character. According to Schopenhauer, the same is true of us.

Schopenhauer distinguishes between three kinds of character in *The World as Will and Representation*. First, there is our "intelligible character" as it exists in the noumenal realm, the will "as thing in itself." It is fundamental, "extra-temporal," "indivisible," "unchanging," and "unfathomable" (1818: 316, 318). Secondly, there is our empirical character, the "appearance" of the will as manifested phenomenally in the "human being's whole pattern of behavior and life history" (1818: 316). This is our character as we more commonly understand it. However, being the mere "appearance" of the more fundamental, primordial, and metaphysically inaccessible "intelligible will," our empirical character is, like reason, completely ineffectual. The empirical character is, as Schopenhauer says, like one of Leibniz's monads, a mere "unfolding" of the will as thing-in-itself (1818: 320). The best we can do, then, is take stock of our intelligible character and organize, to the best of our ability, "the unalterable role of our own person in a thoughtful and methodical manner" (1818: 331). That is, we can learn to avoid or pursue circumstances based on our anticipation of how we will act within them, and if we are able to do so, we will also finally have achieved what Schopenhauer calls an "acquired character."

As some of the above remarks have already anticipated, Schopenhauer held that a person's character—that is, our fundamental or "intelligible" character—is both *innate* and *unchangeable*. "It is not a work of art or accidental circumstances, but the work of

nature itself" (1839: 55). "It remains the same throughout the whole of life" (1839: 51). "Man never changes; as he has acted in one case, so he will always act again—given completely equal circumstances" (1839: 51-2). "People do not change," he says, "rather, their lives and behavior, i.e. their empirical characters, are only the unfolding of their intelligible characters, the development of decided, unalterable dispositions that are already recognizable in the child; thus, behavior is fixed and determined even at birth, and in its essentials stays the same to the very end (1818: 320). Finally, and perhaps most poignantly, "In a word: man does at all times only what he wills, and yet he does this necessarily. But this is due to the fact that he already *is*, there follows of necessity everything that he, at any time, *does*" (1839: 98-9).

3.3 Schopenhauer's Fatalism

As we can see, Schopenhauer's fatalism is motivated in part by the idea that traits of character are "fixed" in the sense that they are (i) innate and immutable. He does not think we have the power to do or to forego doing anything in the future, since a person's reaction to some external stimuli is fated, i.e., non-contingently unavoidable, according to his intelligible character, the "basis" or "ground" through which motives become causally effective. Thus, we lack—unequivocally—internal or "moral" freedom. Not only is the will determined externally by motives or causes, those motives are effective depending on the constitution of my character, the will as thing-in-itself, which I am utterly powerless to change. So, what I do and who I am is necessary and unavoidable. If I wanted to change something about myself, say, become a more honest person, there is no possibility of my becoming so. For, on the one hand, if it were possible for me to become so, I would have already *been* a more honest person. And, on the other hand, whether I am honest in a particular situation depends on the circumstances and motives

that are effective at that time, i.e., on the effect these have on my will as thing-in-itself, not on anything *I* can do to change myself.

Unlike Nietzsche, Schopenhauer also affirms the truth of nomic determinism: "wherever and whenever in the objective, real, material world *anything* changes to a lesser or greater extent, something else must necessarily have changed just before it, and 'something else' before that, and prior to that another, and so on to infinity" (1839: 28). This is because he thinks, following Kant, that all cause and effect relationships, as we perceive them in the phenomenal world, are governed by the "law of causality," a synthetic a priori rule of cognition which holds that, for any change to occur, it must be the effect of some prior cause and follow necessarily as a result (1839: 28). There is thus no room for luck or indeterminacy in Schopenhauer's picture of an agent's future. Unlike CE, the values and moral beliefs I endorse are not the product of contingent or environmental circumstances (e.g., my upbringing or my culture).

In summary, we might say that on Schopenhauer's view the future is *doubly determined*, according to both internal and external causes. From the perspective of the law of causality, all events, including an agent's actions and choices, are necessitated by antecedent causes. From the perspective of intelligible character, one must act as one does according to the influence these external causes have on the will, since we can do nothing to change the constitution of our character, ensuring that we always react to these external stimuli in the manner that we do. As Schopenhauer says:

> To wish that some event had not taken place is a silly self-torture, for this means to wish something absolutely impossible, and is as irrational as is the wish that the sun should rise in the West. Precisely because everything that happens, great or small, happens with strict necessity, it is altogether useless to reflect on how insignificant and accidental were the causes which brought that event about, and how easily they could have been different. For this is illusory. All of them have happened with just as strict necessity and have done this work with just as full a

> power as that in virtue of which the sun rises in the East. We should rather consider the events, as they happen, with the same eye as we consider the printed word which we read, knowing full well that it was there before we read it. (1839: 63-4)[33]

IV. The Development of Nietzsche's Views on Agency and Character

Schopenhauer's fatalistic views of agency and character did indeed leave an indelible mark on Nietzsche, especially early in his career. In *Human, All Too Human* (1878), Nietzsche's first sustained critique of morality, he began espousing a program of complete and "unconditional unfreedom and unaccountability of the will" (HA II/1: 33; see HA II/1: 50), consistently referring to free will as an "error" and "illusion" (HA I: 18, 39, 99, 102, 106), attributing these conclusions to Schopenhauer's "mighty insight" into "the strict necessity of human actions" (HA II/1: 33). There he also argues that we cannot be "accountable" for our nature or character, "inasmuch as it is altogether a necessary consequence and assembled from the elements and influence of things past and present" (HA I: 39). Expanding on this view two aphorisms later, he claims:

> *The unalterable character*. – That the character is unalterable is not in the strict sense true; this favorite proposition means rather no more than that, during the brief lifetime of a man, the effective motives are unable to scratch deeply enough

[33] Curiously, despite our being fated, Schopenhauer still thought we were morally responsible. Following Kant, he believed we possessed noumenal or transcendental freedom, however, his argument for this is not persuasive (see Janaway 2012 for an excellent critical analysis). Briefly, it begins from the observation that we have a "clear and certain" *feeling* of responsibility for what we do, an "unshakeable certainty that we ourselves are the doers of our deeds" (1839: 94). Schopenhauer then supposes there must be a ground or "objective" condition for this feeling (1839: 94). But since our actions are completely fated this condition cannot reside in the ability to *act* otherwise that one did, and so it must instead reside in the person's *character*. We realize a different action "was quite possible and could have happened, *if only he had been another"* (1839: 94). From this observation Schopenhauer draws the questionable inference that "where guilt lies … freedom must also have the same location, namely, in the character of man" (1839: 95). Relying on Kant's distinction between empirical and intelligible character, he infers that the will, as thing-in-itself, possesses an "absolute" and "transcendental freedom" that lies outside of space and time, and as such is not subject to the law of causality (1839: 97). However, even if it were true that we possessed transcendental freedom (and how could anyone know?), it's still unclear how praise and blame would be justified, for it is still the case that these attach to a person's *empirical* character over which we have no control. Praising or blaming another in this case would be like our praising or blaming a videogame character who is merely executing the commands of the person playing the game.

> to erase the imprinted script of many millennia. If one imagines a man of eighty-thousand years, however, one would have in him a character totally alterable: so that an abundance of different individuals would evolve out of him one after the other. The brevity of human life misleads us to many erroneous assertions regarding the qualities of man. (HA I: 41)

As we can see, Nietzsche accepts, in modified form, Schopenhauer's unchangeability thesis. The difference is that Nietzsche offers a naturalistic hypothesis for why traits of character are unchangeable, believing them to be the product of historical circumstances, "the imprinted script of many millennia," and so denies that they are the effect of a noumenal will. He nonetheless agrees that we cannot change our character within "the brief lifetime of a man." Secondly, Nietzsche believes we are "totally unaccountable" for our character because he accepts uncritically Schopenhauer's supposition that a free will must be a *liberum arbitrium indifferentiae*. As I will show here, by the time he wrote *Daybreak* (1881), Nietzsche will have switched courses on both of these issues.

4.2 Character and Nietzsche's Drive Psychology

Though as a practical matter Nietzsche accepts Schopenhauer's conclusion that character is unchangeable in HA, the fact that he takes traits of character to be acquired through naturalistic processes is profoundly at odds with Schopenhauer's view that traits of character are instead the effect of a metaphysical, noumenal will. For it follows from Nietzsche's naturalistic approach, as he already acknowledges above, that character traits are at least *in principle* revisable, since they are themselves the effect of contingent and historical circumstances. Altering a trait of character, on this view, is fundamentally a matter of whether the agent possesses "effective motives" strong enough to modify those drives or dispositions we have acquired as a result of thousands of years of social and cultural selection. Above, Nietzsche expresses skepticism that we can acquire such countervailing motives within the span of one lifetime, and so character is—on *a*

posteriori grounds—unchangeable. So, in HA Nietzsche believes that traits of character are "fixed" in the sense that they are (i) innate and immutable.

Nietzsche will change his mind on the issue of immutability in *Daybreak* three years later; by this point he recognizes that we already possess countervailing motives strong enough to alter our innate moral dispositions, and he also thinks we can develop or acquire such motives through painstaking effort and self-training. That is, Nietzsche will come to endorse the idea that character traits are "fixed" in the sense of (ii): innate and recalcitrant to revision, though in principle revisable. It will be helpful, in order to understand Nietzsche's shift on these issues, to consider his drive psychology which, as we will see, bears many similarities to Schopenhauer's view of agency.

4.2.1 Drives as Innate Dispositions

Schopenhauer, recall, endorses a picture of action in which an act of will is determined by a person's motives in conjunction with her character in the circumstances, or: *A is determined by M in conjunction with C in X.* Nietzsche modifies this account in a few crucial respects, but in its basic outlines it remains true for him as well. The first and main difference is that Nietzsche replaces talk of motives with talk of *drives* [*Trieb*]. Nietzsche, anticipating Freud, took our subconscious mental life to be far more important to explaining our actions and beliefs than our conscious beliefs and motives. He not only believes that the desires and motives which reveal themselves in consciousness are the effect of drives, he believes that all actions, including those we consciously will, are determined by drives (BGE 19). Moreover, he believes that the self is nothing but the "social structure of the drives and affects" (BGE 12; see Anderson 2012, Clark and Dudrick 2012), and that our character is constituted by the "order of rank the innermost drives of [one's] nature stand in relation to each other" (BGE 6).

Also, like Schopenhauer's conception of a motive, Nietzsche thinks drives cause, or at least apt to cause, an agent to act in some particular manner. Specifically, Nietzsche conceives of drives as plastic dispositions to act in terms of the characteristic goal or activity they have been naturally or culturally selected to induce the agent to perform (Richardson 2004). For instance, the drive to eat is a disposition to engage in the activities necessary to nourish the body with food, though it may be satisfied in ways other than digestion (e.g., if one is on a diet and hungry, the drive to eat may be put in abeyance by chewing gum). As Paul Katsafanas has argued (2016: 106), a drive induces behavior by producing within the agent an *affective orientation* that colors her perception of her environment and makes attractive the activities and objects that would discharge and satisfy the drive; in this way, drives "structure" the agent's perceptions, affects, and reflective thought.[34] For example, if a person is hungry, features of her environment that remind her of food will become more salient and grab her attention until she eats (e.g., she will be extra sensitive to the smell of her colleague's lunch, or the distant beeping of the microwave in the breakroom).

Since drives are the products of natural or cultural selection, it makes sense that Nietzsche would think that traits of character are innate. As noted previously, a person's character is a social identity, consisting of evaluative descriptions that attach to enduring dispositions exhibited in the agent's actions and beliefs. Nietzsche provides an answer as to how we acquired these dispositions in *The Gay Science*:

> *Herd instinct.* – Wherever we encounter a morality, we find an evaluation and ranking of human drives and actions. These evaluations and rankings are always

[34] Similarly, Maudemarie Clark and David Dudrick claim: "Nietzsche's view seems to be that all of the drives have their own point of view on the world in the sense that a drive, when active, turns the spotlight of one's cognitive capacities on those features of reality which will increase the drive's chances of attaining its end" (2012: 145).

> the expression of the ends of a community and herd: that which benefits *it* the most – and second most, and third most – it is also the highest standard of value for all individuals. With morality the individual is instructed to be a function of the herd and to ascribe value to himself only as a function. Since the conditions for preserving one community have been very different from those of another community, there have been very different moralities; and in view of essential changes in herds and communities, states and societies that are yet to come, one can prophesy that there will yet be very divergent moralities. Morality is herd-instinct in the individual. (GS 116)

As revealed above, we acquired the moral dispositions that constitute our character essentially by being trained to do so over "many millennia" (HA I: 41). Descriptively, the function of morality is to "rank" human actions and drives so as to encourage those behaviors and dispositions that allow the community to flourish, and to discourage those that would lead to its dissolution (see GM II: 3). "Morality," in the sense described above, refers to the disposition to act on these pro-social "herd instincts" that are well-adapted to a particular environment. As we internalize these dispositions over time, we acquire a natural "rank order" of the drives in which these dispositions are given deliberative priority whenever the agent faces conflicts about what to do. Thus, on Nietzsche's view, we are born with the moral dispositions that constitute our character, which then come to be reinforced and refined through culture and upbringing.

4.2.2 Self-Opacity and Mutability

Nietzsche never changes his mind about drives being innate, but this implies neither that such traits are unchangeable, nor that we lack free will because we are born with dispositions to behave or think in ways that are reinforced through our upbringing and culture. More problematic for free will is the fact that, following Schopenhauer, Nietzsche conceives of consciousness as little more than a veneer that masks the underlying causes of our actions (the drives). "However far a man may go in self-knowledge," he writes in *Daybreak*, "nothing however can be more incomplete than his

image of the totality of *drives* which constitute his being. He can scarcely name even the cruder ones: their number and strength, their ebb and flow, their play and counter-play among one another" (D 119). Here there are two problems. First, we lack reliable epistemic access to our subconscious drives, and for that reason our real motives remain *opaque* to us. Secondly, when we reflect on our motives as they are represented consciously, we do so in terms of linguistically articulable folk-psychological concepts that effectively guarantee that "their real nature remains introspectively inaccessible," according to Mattia Riccardi (2015: 238; see GS 354; D 116, 129). That is, when we attempt to rationalize our behavior or reflect on our conscious motives, because we must do so in terms of socially constructed concepts, our doing so is so incomplete and superficial that it amounts to a "falsification" (GS 354) of the drives themselves. For this reason, every action is "unknowable" (GS 335).

Also, like Hume and Schopenhauer, Nietzsche believes that reason is effective only as a means to achieving something else we already desire: it merely does the bidding of the agent's subconscious drives (see D 119, 120; GS 335; BGE 187; TI VI: 3). As Maudemarie Clark and David Dudrick claim, drives "monopolize the person's cognitive capacities in the service of their own ends" (2012: 146). This is confirmed in an important aphorism titled "self-mastery," where Nietzsche discusses "six different ways to combat the vehemence of a drive" (D 109). (We'll discuss these six ways in a moment.) He concludes the passage by noting:

> … *that* one *desires* to combat the vehemence of a drive at all, however, does not stand within our own power; nor does the choice of any particular method; nor does the success or failure of this method. What is clearly the case is that in this entire procedure our intellect is only the blind instrument of *another drive* which is a *rival* to the drive whose vehemence is tormenting us … While 'we' believe we are complaining about the vehemence of a drive, at bottom it is one drive *which is complaining about another*; that is to say: for us to become aware that we are

> suffering from the *vehemence* of a drive presupposes the existence of another equally vehement or even more vehement drive, and that a *struggle* is in prospect in which our intellect is going to have to take sides. (D 109)

Nietzsche's theory of agency, because it takes the drives to be innate, opaque, and, to a considerable extent, beyond our ability to control, does indeed represent a serious threat to free will and moral responsibility. Yet, as we can see in this passage from *Daybreak*, by this time Nietzsche does *not* think that a person's drives are immutable.

In fact, what Nietzsche says above is rather a straightforward implication of his naturalistic approach, modifying the earlier view in *Human, All Too Human* in one crucial respect: Nietzsche now recognizes that one need not live "eighty-thousand years" to develop conflicting drives in order to alter one's character, as constituted by the natural rank order of one's innermost drives. Instead, to alter a trait of character, it must be the case that the agent possesses an "equally vehement" drive, or a coalition of such drives, which generates conflict. This *alteration principle*, as I will call it, establishes both a *constraint* and confers a type of freedom relevant to self-constitution.

First, I am not able to alter a drive simply by deciding to do so or through an "act of will," understood as an ability to act that is not itself the effect of some other drive(s). For it to even be *possible* to alter a drive, and whether I am able to alter it successfully, is in this sense not up to me. The ability to alter a drive instead depends on my *other* drives, on the agent's self as currently constituted, and specifically on the condition whether an "equally vehement drive," or a coalition of such drives, is strong enough subdue it. Secondly, though, we have the ability to alter the *drives themselves*, including those that constitute our character, our "innermost nature" (BGE 6), presuming the first

condition is satisfied. Crucially, this means we have an ability to change the constitution of the self that Schopenhauer and Leiter deny we have.[35]

4.3 "What we are at liberty to do"

Nietzsche is clearest in his rejection of Schopenhauer's unchangeability thesis, as I have already stated, in *Daybreak* (1881):

> *What we are at liberty to do.* – One can dispose of one's drives like a gardener and, though few know it, cultivate the shoots of anger, pity, curiosity, vanity as productively and profitably as a beautiful fruit tree on a trellis; one can do it with the good or bad taste of a gardener and, as it were, in the French or English or Dutch or Chinese fashion; one can also let nature rule and only attend to a little embellishment and tidying-up here and there; one can, finally, without paying any attention to them at all, let the plants grow up and fight their fight among themselves—indeed, one can take delight in such a wilderness, and desire precisely this delight, though it gives one some trouble, too. All this we are liberty to do: but how many know we are at liberty to do it? Do the majority not *believe* in *themselves* as in complete *fully-developed facts?* Have the great philosophers put their seal on this prejudice with the doctrine of the unchangeability of character? (D 560; see D 364, 382; GS 296)

That this passage is meant as a refutation of Schopenhauer cannot be doubted, especially when one considers that all of the evidence he cites in favor of the unchangeability of character throughout *On the Freedom of the Will* is anecdotal.[36]

But what is the nature of the disagreement? According to Leiter, the disagreement is merely *apparent*. He surmises that Nietzsche misunderstood

[35] Leiter takes it to be an implication of D 109 that the self is "merely the arena in which the struggle of the drives plays itself out" (2002: 103). *One* view Nietzsche holds about the self would imply this, namely, when he equates the self with the "body" (Z I: 4), but he is equally committed to the self as a *normative ideal*, a self we can *create* by shaping the disparate elements that have simply been given to us into a unified whole for which we can take credit (GS 290, 335; HA II: 366). This, not incidentally, is also how we realize our freedom: "For what is freedom? That one has the will to assume responsibility for oneself" (TI IX: 38). We are "free" for Nietzsche to the extent that there is conflict that can be overcome, or to the extent that one is a *conflicted multiplicity* that can be made unified, as opposed to a predetermined simplicity.

[36] As Janaway observes, Schopenhauer derives his evidence from "popular sayings and attitudes, some from poets and dramatists, some from authorities in classical antiquity – though in truth this really only establishes that it has often been believed that character is individual, inborn, and unchanging" (2012: 442-3; see Schopenhauer 1839: 49-60).

Schopenhauer's view, suggesting that he failed to recognize that what he says above is just the same thing Schopenhauer meant by "acquired character." "Yet it appears this is precisely Schopenhauer's view as well (though Nietzsche seems not to have recognized it): the 'unalterability' of character does not, it seems, entail that there is no 'gardening' work to be done on the basic ingredients (e.g., the drives) which constitute the 'character'" (Leiter 2002: 62-3). In fact, however, it appears Leiter has misunderstood *Schopenhauer's* view, for he does hold that it is futile to try and change the dispositions which constitute our character:

> But no moral influence can reach further than correction of cognition, and the undertaking to remove the failings of character of a man by means of talk and moralizing and thus to reform his character itself, his essential morality, is exactly like the attempt to change lead into gold by external action, or by means of careful cultivation to make an oak produce apricots. (Schopenhauer 1839: 54)

Acquired character, as noted previously, is concerned exclusively with *external modification* of one's circumstances, not with altering the basic internal structure of the self. For example, if I am a coward—if it is typical of me to run or flee in the face of conflict—per Schopenhauer's view of character I can never change this about myself. If I am confronted by a threat, in normal circumstances I will always flee and never fight.[37] The best I can do is avoid the *circumstances* in which my cowardliness would manifest itself. So, I might, for example, avoid the playground after school or the locker room after PE class, since I know bullies frequent both of these places. However, I cannot

[37] Recall, it is still the case that circumstances play a role in determining action on Schopenhauer's view. So, if we are to imagine that a bear is charging my daughter, I might be able to muster up the courage *in this circumstance* to act bravely.

change my cowardliness, and so if I am confronted by a bully, I will inevitably cower in the face of fear. "Acquired character" is limited to external alteration of circumstances.[38]

Nietzsche clearly means to be saying something different in D 506. He does not disagree that the drives (or "plants") are innate. We might suppose that the "shoots of anger, pity, curiosity, vanity" are part of us, and necessarily so (for if they were not, we would not be who we *are*). He moreover does not deny that we are constrained by these, if they are constitutive of our character, our "innermost drives" (BGE 6). The difference is that he thinks we can alter the drives themselves:

> First, one can avoid opportunities for gratification of the drive, and through long and ever longer periods of non-gratification weaken it and make it wither away. Then, one can impose upon oneself strict regularity in its gratification: by thus imposing a rule upon the drive itself and enclosing its ebb and flood within firm time-boundaries, one has then gained intervals during which one is no longer troubled by it—and from there one can perhaps go over to the first method. Thirdly, one can deliberately give oneself over to the wild and unrestrained gratification of a drive in order to generate disgust with it and with disgust to acquire a power over the drive … Fourthly, there is the intellectual artifice of associating its gratification in general so firmly with some very painful thought that, after a little practice, the thought of its gratification is itself at once felt as very painful … Fifthly, one [can] .. deliberately [subject] oneself to a new stimulus and pleasure and thus [direct] one's thoughts and plays of physical forces into other channels … Finally, sixth: he who can endure it and finds it reasonable to weaken and depress his *entire* bodily and physical organization will naturally thereby also attain the goal of weakening and individual violent drive. (D 109)

None of the methods described above presume that we can rid ourselves of a drive merely by deciding to do. They also do not rule out the possibility that we may not be able to subdue a vehement drive. Again, according to Nietzsche's alteration principle, doing so will only be possible and successful if there is a causal link between the unruly

[38] "If we clearly recognize our failings and weaknesses as well as our good qualities and strengths, once and for all, if we plot our goal accordingly and accept what we cannot do; then this is the surest way of escaping (as far as possible, given our individuality) the bitterest of all sufferings" (1818: 333).

drive and another drive (or drives) which are powerful enough to subdue it. And that fact is not up to us.[39]

Let's consider Nietzsche's alteration principle at work. Here, as before, let's imagine I am cowardly, but let's also imagine that I am able to associate my drive to flee when confronted by a bully with a "painful thought" in order to subdue it. For this to be possible, I must possess an "equally vehement drive" (D 109), and so let's imagine that I also care that others do not perceive me as a coward. Let's stipulate that I also care about my reputation and masculinity. If this is the case, I can train myself to be less cowardly by, e.g., associating the drive to flee with the mockery of my classmates. If I continually associate the two thoughts—the mockery of my classmates and my strong inclination to flee—I can succeed in making the latter disposition uncomfortable enough to overcome it, so long as these other drives (the desire for a good reputation, to be perceived as masculine) are stronger. Sure, whether I am *successful* in doing so will depend on the strength of these competing drives (and there would never *be* a conflict in the first place if I did not have these competing drives). But what is significant is that, for Nietzsche, the presence of conflict is an enabling condition alteration: through *conflict* we make a causal difference to the constitution of the self; we realize an ability Schopenhauer denied we possessed.[40]

So, the disagreement between the two thinkers is centered around a kind of *freedom* or "liberty" Nietzsche thinks we have and which Schopenhauer denies: the

---

[39] So, in other words, some drives *may* be (i) innate and immutable. The important thing is that Nietzsche does not think a person's *character* consists entirely of such drives.

[40] Also, since Nietzsche allows for environment and circumstances to play a role in shaping character, it would seem that, in principle at least, *every* drive is alterable, so long as "equally vehement" drives can be acquired or cultivated. As Anderson (2012: 213) and Richardson (1996: 44-52) note, drives can be combined to create new drives.

ability to alter the composition of our character. Schopenhauer denied that we possess this freedom because he was a *source incompatibilist*; he believed that to be the causal source of one's action, the agent's will must be a *liberum arbitrium indifferentiae*, or that she must possess uninhibited *leeway freedom*. As we have seen, in *Human, All Too Human* Nietzsche agreed, but in *Daybreak* he began endorsing a *source compatibilist* conception of freedom: we are "at liberty" to alter the constitution of our character, within the constraints established by the natural rank order of our drives, according to Nietzsche's alteration principle. As Frankfurt notes,

> The question of whether the person is responsible for his own *character* has to do with whether he has *taken responsibility for* his characteristics. It concerns whether the dispositions at issue, regardless of whether their *existence* is due to the person's own initiative and causal agency or not, are characteristics with which he identifies and which he thus by his own will incorporates into himself as constitutive of what he is (1987: 171-2).

Like Frankfurt, Nietzsche denies that we possess *leeway freedom*—the ability to do otherwise. We are not responsible for the drives we in fact have; we do not have the ability to have a character other than the one we have. We have the ability to shape and "give style" (GS 290) to the character we've been given, thus making it our own. Character, then, is not unchangeable, not "fixed" in the sense of (i) being innate and immutable, precisely because we are "free" in this sense—*that* is the crux of their disagreement.

Thus, Nietzsche is a "fatalist" in Leiter's sense—he holds that our future life trajectories are constrained, perhaps even significantly so, by the natural rank order of our drives. He is also a fatalist in the sense of ISE: the self-as-constituted *must* be the way that it is, simply because if it weren't it would not *be* what it is. However, while both forms of "fatalism" rule out or possessing *leeway freedom*, neither is incompatible with

our possessing *source freedom*, since Nietzsche holds that we have the freedom or ability to alter this rank ordering, given these constraints.[41] To see why this is, let's now consider Nietzsche's compatibilist conception of source freedom in more detail, as well as its implications for free will and responsibility.

V. Free Will and Responsibility

Free will is commonly regarded as a condition that must be satisfied in order for ascriptions of responsibility to be justified. This is how Leiter proposes Nietzsche understands free will (2002: 87), and he also maintains that, like Schopenhauer, Nietzsche is an incompatibilist. Accordingly, since Nietzsche's fatalism on Leiter's view implies that one's choices and actions are determined by the combined influence of type-facts plus environment and circumstances, he takes Nietzsche to be a so-called "hard determinist," or one who affirms the truth of determinism (fatalism) and denies that we have free will, and thus are not morally responsible. As we just saw, however, Nietzsche recognizes and is committed to our possessing a type of freedom that is compatible with our being fated, and so Nietzsche's fatalism does not seem to have the incompatibilist implications that Leiter contends (2002: 88-91).

As I argue here, Nietzsche's fatalism only shows that we lack *libertarian* free will, and thus are not "ultimately responsible" (BGE 21) for ourselves or our actions. Leiter mistakenly interprets this passage in which Nietzsche criticizes libertarian, or "causa sui" (BGE 21) free will, as a necessary condition for moral responsibility, when in fact Nietzsche is not concerned with *ascribing* responsibility at all in this passage. Instead,

[41] But doesn't this imply that, when we successfully alter a drive, we bring into existence an alternative possibility, thus redirecting the course of fate? No, the agent's contribution/effort introduces only temporal variation, not variation across possibilities. We make a difference to *how* fate unfolds necessarily, one might say, but we make no difference to the course it takes.

the passage relies on Nietzsche's positive, source compatibilist conception of freedom, which is a condition of *taking* responsibility for oneself. As we will see, Nietzschean freedom is realized by taking ownership of the disparate and conflicting elements of the self we are fated to have and shaping them into a unified whole: "For what is freedom? That one has the will to assume responsibility for oneself" (TI IX: 38).

5.1 The Causa Sui

That Nietzsche means to reject libertarian free will is uncontroversial. Following Robert Kane, we will define libertarian free will as "the power of agents to be the ultimate creators (or originators) and sustainers of their ends or purposes" (1998: 4). In order for this to be the case, Nietzsche thinks the will must be a *causa sui*, or self-cause (BGE 21), but, like Schopenhauer, he believed free will in this sense to be completely illusory:

> The *causa sui* is the best self-contradiction that has been conceived so far, it is a sort of rape and perversion of logic; but the extravagant pride of man has managed to entangle itself profoundly and frightfully with just this nonsense. The desire for "freedom of the will" in the superlative metaphysical sense … the desire to bear the entire and ultimate responsibility for one's actions oneself, and to absolve God, the world, ancestors, chance, and society involves nothing less than to be precisely this *causa sui* and … to pull oneself up into existence by the hair, out of the swamps of nothingness. (BGE 21)

In order to be a causa sui, an agent's choice or action must arise spontaneously through the efforts of her will alone and the will cannot be determined by any prior influence. Above, Nietzsche specifies a conception of responsibility, "ultimate responsibility," that depends on the causa sui. If one is ultimately responsible, then the person's actions and choices must be immune to luck, which the causa sui, since it understands the agent to be the ultimate, buck-stopping source of her choices and actions, secures.[42] The causa

[42] The causa sui is similar to being an "agent-cause." As Roderick Chisholm claims, "If we are responsible, and if what I have been trying to say is true, then we have a prerogative which some would attribute only to God: each of us, when we act, is a prime mover unmoved. In doing

sui thus amounts to possessing free will "in the superlative metaphysical sense" (BGE 21), because it implies that, whenever a person makes a choice or performs some action, it is *always possible* for her to choose or act otherwise than she in fact does.

It is a straightforward and unproblematic implication of Nietzsche's fatalism that we lack libertarian free will and do not bear ultimate responsibility for ourselves. First, we have already seen that a person's actions and choices are determined by her subconscious drives, and so it would be impossible for us to have a will that was undetermined by any prior cause. Secondly, a person cannot bear ultimate responsibility for herself because her character is circumscribed by the innate constitution of her drives as well as her environment and circumstances. Consequently, as Leiter has argued, every choice and action we make is "causally determined by something about the 'way we already are' – including those operations of will in which we attempt to alter the 'way we already are'" (2002: 90). Indeed, as we have seen, Nietzsche's alteration principle preserves this lesson, since a person can only change an inherent drive or disposition if she possesses an "equally vehement" conflicting drive or a coalition of such drives (D 109), and so we can only change the constitution of the self on the basis of what we *already* are.

Yet, it does not follow from these observations that Nietzsche thinks we must be a causa sui to possess free will. For that to be the case, he would have to agree with Schopenhauer that a free will must be uncaused, but he nowhere makes that assertion in the above passage. In fact, he asks us to "put [the causa sui] out of [our] head[s] altogether," along with the idea of "unfree will" (BGE 21). Elaborating, he claims:

---

what we do, we cause certain events to happen, and nothing—or no one—causes us to cause those events to happen (1964: 34).

> The "unfree will" is a mythology; in real life it is only a matter of *strong* and *weak* wills. It is almost always a symptom of what is lacking in himself when a thinker senses in every "causal connection" and "psychological necessity" something of constraint, need, compulsion to obey, pressure, and unfreedom; it is suspicious to have such feelings—the person betrays himself. And, in general, if I have observed correctly, the "unfreedom of the will" is regarded as a problem from two entirely opposite standpoints, but always in a profoundly *personal* manner: some will not give up their "responsibility," their belief in *themselves*, the personal right to *their* merits at any price (the vain races belong to this class). Others, on the contrary, do not wish to be answerable for anything, or blamed for anything, and owing to an inward self-contempt, seek to *lay blame for themselves somewhere else*. (BGE 21)

So, quite apart from drawing the conclusion that one must be a causa sui to possess free will, Nietzsche instead draws a *psychological* lesson about those who think this is true. Both groups of people think this is true based on their unwillingness to *take* responsibility for themselves unless one is a causa sui. Proud individuals, because they want to take complete credit for everything they've accomplished, claim to have causa sui free will. Those who are dissatisfied with themselves, on the other hand, are unwilling to take any responsibility for who they've become, and so they claim no one is a causa sui.

So, I think Leiter is mistaken to interpret the above passage as articulating what Nietzsche takes to be a condition of moral responsibility.[43] In fact, the passage is not concerned with free will as a condition of *ascribing* responsibility at all. Nietzsche is instead interested above in freedom of the will as a condition of *taking* responsibility. To understand this connection, we will have to consider Nietzsche's positive conception of freedom. This will also help us make sense of his puzzling claim that "in real life it is only a matter of *strong* and *weak* wills" (BGE 21).

[43] He takes Nietzsche's point in this passage to be that "Our 'will' is an artifact of the facts about us, and thus cannot be the source of genuinely autonomous action (the sort that would ground responsibility)" (Leiter 2002: 89).

5.2 Autonomy as a Self-Relation

Much of what I have to say here about Nietzsche's positive conception of freedom has been covered in greater detail by other scholars. It is now widely accepted that Nietzsche defends a positive, source compatibilist conception of freedom as autonomy and as a normative ideal.[44] According to this conception, a person is free or autonomous to the extent that one's actions are self-governed or one's self is deliberately ordered and unified (see GS 98, 99, 290, 335; GM II: 2; TI IX: 38, 49). As this suggests, freedom for Nietzsche is not an all or nothing affair: we can be more or less autonomous, either with respect to the freedom of action, or with respect to the self. In fact, Nietzsche takes the former to be dependent on the latter. As Lanier Anderson has argued, Nietzsche conceives of autonomy as a "distinctive form of self-relation" achieved "through a kind of self-creation" (2013: 456). We will discuss these two ideas in turn.

Nietzsche flips traditional approaches to freedom on their heads. Typically, the presence of conflict and constraint, whether internal or external, is taken to be an impediment to realizing one's freedom. For Nietzsche, these are instead conditions of our having freedom at all. We are not free merely in virtue of having certain capacities, say, a capacity for reflective endorsement, freedom must be *earned*; it is "something one *wants*, something one *conquers*" (TI IX: 38). "How is freedom to be measured in individuals and peoples? According to the resistance which must be overcome, according to the exertion required, to remain on top. The highest type of free men should be sought where the highest resistance is constantly overcome: five steps from tyranny, close to the threshold of the danger of servitude" (TI IX: 38).

[44] See Anderson (2012, 2013), Gemes (2006), Janaway (2006, 2007), Katsafanas (2016), Richardson (2004), Ridley (2007), though he argues Nietzsche was not a "compatibilist," and Rutherford (2011). Clark (2015) additionally argues that Nietzsche is committed to a compatibilist conception of freedom that would ground ascriptions of moral responsibility.

These remarks are curious, to say the least. They would seem to imply that the slave who overcomes his bondage after many years of servitude would be more "free" than the person born of privilege and with all the opportunities in the world at her disposal. Surely the latter had many opportunities available to her that the former did not, and certainly the former was constrained in ways that would count against his being "free" at all. These considerations, though, hinge on whether these individuals have the *external* or *leeway* freedom to do what they want. Like Schopenhauer, Nietzsche was more interested in freedom *of* the will, or whether the agent is *internally* free to act as she wants; whether she can be the *source* of her action.

More precisely, he was concerned with willing as an *activity* that involves a conscious representation of some goal constitutively pursued against resistance, i.e., as a kind of striving that is related essentially to constraint.[45] "That which is termed 'freedom of the will'," he tells us, is "essentially the affect of superiority in relation to him who must obey: 'I am free, 'he' must 'obey'—this consciousness is inherent in every will" (BGE 19). "'Freedom of the will'—that is the expression for the complex state of delight of the person exercising volition, who commands and the same time identifies himself with the executor of the order—who, as such, enjoys also the triumph over obstacles"

[45] Nietzsche analysis of willing bears many similarities to Schopenhauer's (see Janaway 2007, Reginster 2006). However, Nietzsche denies that there is a *faculty* of will (BGE 19, A 14), so he is not concerned, as Schopenhauer was, with the internal conflict between *the will* and a motive. He takes willing to be "something *complicated*, something that is a unit only as a word" (BGE 19), which involves three distinct phenomenological components: (i) an initial and often imperceptible feeling of attraction toward or aversion away from some object, along with accompanying "muscular" or bodily feelings; (ii) "thinking," specifically a "ruling thought," a conscious representation of the object, motive, or the course of action to be pursued; (iii) the "affect of command," "the inward certainty that obedience will be rendered" (BGE 19). Clark and Dudrick (2012: 178-81) argue, contra Leiter (2007), that Nietzsche's paradigm cases of acts of will are not of voluntary actions or deliberate decisions, but instead cases of *willpower*, a "situation of psychic conflict and struggle in which a person is faced with a choice between alternatives" (2012: 181). I concur, for reasons I cannot get into here. It is more important for my purposes that we recognize, again, that Nietzsche comes to reject one of Schopenhauer's basic assumptions.

(BGE 19). Importantly, as this reveals, Nietzsche, unlike Schopenhauer, does not think a "free" will must be absolutely and negatively free, a *liberum arbitrium indifferentiae*; freedom is instead something that is realized by *overcoming* constraint or resistance.

Anderson (2013) proposes that we understand Nietzsche's positive conception of freedom in terms of Nietzsche's curious claim that "The 'unfree will' is a mythology; in real life it is only a matter of *strong* and *weak* wills" (BGE 21). In *Twilight*, Nietzsche tell us that the "essential characteristic of a "strong will" is precisely the ability "*not* to will—to *be able* to suspend decision" (TI VIII: 6); by contrast, weakness of will is "the inability *not* to respond to a stimulus" (TI V: 2). So a strong will can resist, say, the compulsion to lash out at someone who has offended you, or the temptation to eat another chocolate, whereas a weak will succumbs to these. Why is that? According to Anderson, whether one has a strong or a weak will is incumbent upon the underlying constitution of the self:

> Weakness amounts to a characteristic form of inner division that makes us vulnerable to being pushed around by our drives—and pulled around by external stimuli, because the drives and affects responsive to those stimuli are insufficiently integrated with the rest of our attitudes, and so elude the kind of control by the whole self that would enable us to resist the stimuli. Strength amounts to the converse form of inner unity, affording an integrated self that can control its constituent drives and so has the ability '*not* to will' even in cases where some drive is demanding it. (2013: 456)

Thus, the "strong" will is free or autonomous *because* the self is well-ordered and unified, and so able to resist deviant drives, and the "weak" will is unfree (psychologically) *because* it is not integrated, and so compulsively reacts to drives. This is what Anderson means when he says, for Nietzsche, autonomy involves a kind of "self-relation." The free or autonomous agent is self-controlled or self-governing, where this is achieved by having a highly integrated self, and so Nietzsche's idea of free action comes to depend on his idea of freedom as it relates to the person.

5.3 Fatalism and Self-Creation

Of course, as we have seen, this self that is created and a normative ideal for Nietzsche is not created *ex nihilo*. It is constructed from innate raw materials, the drives and dispositions one is born with, which are refined and reinforced through culture and upbringing. This is the self we are *fated* to have, which Nietzsche equates with "the body" more generally.[46] "Behind your thoughts and feelings, my brother, stands a mighty ruler, an unknown sage—whose name is self. In your body he dwells; he is your body" (Z I: 4). "Your self laughs at your ego and at its bold leaps. 'What are these great leaps and flights of thought to me?' it says to itself. 'A detour to my end. I am the leading strings of the ego and the prompter of its concepts'" (Z I: 4). To say that the self is the body is to stress the point that the self is *embodied*, consisting of the person's drives, the "imprinted script of many millennia" (HA I: 41), and not simply a container held together by one's skin, skeletal structure, and muscles. Moreover, a person's drives are at least loosely unified insofar as they are dispositions that are well-adapted to a particular environment and way of life. But this unity alone is not enough to make the agent autonomous or self-governing.

Becoming "free" is a *task* that each person must undergo by assessing those innate dispositions and values one has accepted uncritically, and insofar as she is capable, creating a self for which one can take ownership out of these very elements. As Nietzsche tells us, *"everyone who wishes to become free must become free through his own endeavor* ... freedom does not fall into any man's lap as a miraculous gift" (UM IV: 11, GS 99). Elsewhere Nietzsche describes this task in terms of "giving style" to one's innate character:

[46] Anderson (2012) refers to the idea of a Nietzschean "minimal self" in a similar capacity.

> *One thing is needful.*- To 'give style' to one's character – a great and rare art! It is practiced by those who survey all the strengths and weaknesses that their nature has to offer and then fit them into an artistic plan until each appears as art and reason and even weaknesses delight the eye. Here a mass of second nature has been added; there a piece of first nature removed – both times through long practice and daily work at it. Here the ugly that could not be removed is concealed; there it is reinterpreted into sublimity. Much that is vague and resisted shaping has been saved and employed for distant views ... In the end, when the work is complete, it becomes clear how it was the force of a single taste that ruled and shaped everything great and small. (GS 290)

It is important to recognize that Nietzsche takes the kind of "self-creation" at consideration here to be *constrained*, and perhaps in some cases significantly so. Some pieces of our "first nature" can be removed entirely; some bits of "second nature" can be added anew. But some elements, "the ugly," cannot be changed. They *must* remain. However, they can still be "concealed" (i.e., successfully repressed) or "reinterpreted into sublimity" (i.e., sublimated).

Constraint is again emphasized by Nietzsche in another important passage where he talks about self-creation:

> Your judgment, 'that is right' has a prehistory in your drives, inclinations, aversions, experiences ... *that* you hear this or that judgment as the worlds of conscience, i.e., *that* you feel something to be right may have its cause in your never having thought much about yourself and in your blindly having accepted what has been labelled *right* since your childhood ... Your insight into *how such things as moral judgments could ever have come into existence* would spoil these emotional words for you ... Let us therefore *limit* ourselves to the purification of our opinions and value judgments and to the *creation of tables of what is good that are new and all our own*. (GS 335)

However, near the end of this passage he makes it even clearer that the self which is "created" is instead only a version of what one *already is*: "We, however, want to *become who we are*–human beings who are new, unique, incomparable, who give themselves laws, who create themselves!" Again, we see echoes of Schopenhauer in this view (1839: 98-9), but for Nietzsche our already being a certain way does not preclude

the possibility of our becoming something different, even if what we become is only a more perfected and refined version of what we already *are* (and must be, per ISE).

We are now in a position to make sense of Nietzsche's conception of freedom: "For what is freedom? That one has the will to assume responsibility for oneself" (TI IX: 38). A free individual *takes* responsibility for the self she has been "fated" to have. To do this she must, first, meticulously assess her inner drives. This is no easy task, since the drives are by nature opaque to us and self-deception too easy. It requires that we use all available resources at our disposal: psychology and science, the often uncomfortable feedback of those close to us, and scrupulous self-reflection. No less difficult, we must also reflect on and reject or endorse those values which we have merely been conditioned to accept. This, like the first task, will take an entire lifetime and there is no sense in which it can be said to be "completed." "To become what one is," Nietzsche tells us in *Ecce Homo*, "one must not have the faintest notion *what* one is" (EH II: 9).[47] That is, becoming a self is a task that can only begin in self-ignorance, and proceed at every stage through overcoming self-ignorance. It is only by overcoming self-ignorance on Nietzsche's view that we can take a *stand* for anything, to be responsible for who we are, and to the extent that we achieve this, be autonomous or self-governing. In other words, in order to be free we must "will a self."

> Will a self. – Active, successful natures act, not according to the dictum 'know thyself,' but as if there hovered before them the commandment: *will* a self and thou shalt *become* a self. – Fate seems to have left the choice still up to them;

[47] Curiously, Nietzsche considers it prudent to ignore the imperative "know thyself" in the process of becoming a self, since it tends to make one "smaller, narrower, mediocre" (EH II: 9). To be clear, he is not inveighing against self-knowledge here, but self-knowledge come too soon, and thus the imperative to know thyself before one's inner nature has been sufficiently revealed. One "must be kept clear of all great imperatives" (EH II: 9), he tells us, because one must be willing to take on a variety of perspectives through this process of self-discovery. "From this point of view even the *blunders* of life have their own meaning and value" (EH II: 9), for it is precisely through such "wrong roads" that we learn the most about ourselves.

> whereas the inactive and contemplative cogitate on what they *have* already chosen, on *one* occasion, when they entered into life (HA II: 366).

VI. Conclusion

In this chapter, I have been concerned with critically assessing Nietzsche's fatalism and its implications for free will and responsibility. We saw that Nietzsche's fatalistic pronouncements actually rest on two sets of considerations, those surrounding identity (ISE) and those surrounding constraint (CE). My focus has been on CE, since Leiter takes it to be an implication of our being "fated" to have a self, as constituted by the innate drives and dispositions given to us by natural and cultural selection and reinforced through upbringing, that we lack free will. Had Nietzsche maintained, as Schopenhauer did, that a free will must be absolutely and negatively free, and also that we can do nothing to alter the innermost drives and dispositions that comprise one's character, it would be an implication of Nietzsche's fatalism that we lack free will. However, as we have seen, Nietzsche's fatalism implies only that we lack "causa sui" free will, and thus do not bear "ultimate responsibility" (BGE 21) for ourselves or our actions, which does not imply that we lack free will altogether and bear no responsibility for ourselves. More precisely, Nietzsche is an incompatibilist about leeway freedom, but he is a compatibilist about source freedom.

Nietzsche's positive conception of source freedom is realized by *overcoming* constraint, precisely by "creating" a self from the raw materials (the drives) we have been given. This is accomplished according to Nietzsche's alteration principle, which holds that in order to alter a drive or disposition there must exist an "equally vehement" (D 109) drive, or coalition of such drives. Consequently, it is not completely "up to us" whether we have the tools necessary to become a self, but to the extent that one is a

conflicted multiplicity that can be made unified, one is able to become autonomous and free. This is a kind of internal freedom that Schopenhauer denied we possessed, and on Leiter's view we lack it too. For Schopenhauer we are like water, which can only move once acted upon from without by an external cause, and for Leiter, similarly, we are like a tomato plant that cannot help but grow tomatoes so long as our environment is hospitable. It should now be clear how both of these analogies err, as applied to Nietzsche. Water and tomato plants, in comparison to human beings, are *simple* structures—not merely chemically, but *psychologically*. Leiter is right to stress that, on Nietzsche's view, a substance's future trajectories are *constrained* by what one already is. Water cannot boil at 80 degrees Fahrenheit, just as a tomato plant cannot grow apples, but water and tomato plants do not have competing dispositions to do such things. Lacking conflict, they lack the possibility to alter their fundamental constitution, merely in virtue of being simple.

A better analogy is the one Nietzsche himself offers. The human self is a "garden" (D 506) in which many native plants (drives) compete to grow (be causally effective). Depending on the size and breadth of our garden's plot, and the ability to "cross-pollinize," as it were, we are capable of growing corn, tomatoes, and many other plants in between. Indeed, the more plants we have competing to grow, the more one can make a garden that is truly unique and our own. However, we are not at liberty to choose the size of our garden's plot, the plants that are native to it, or the climate they grow in. These are all *fated*, though our being "fated" in this sense does not preclude our being responsible for what takes root.

References

Works by Nietzsche

BGE: *Beyond Good and Evil*. 1886. In W. Kaufmann, trans. and ed., *The Basic Writings of Nietzsche*. New York: Modern Library Edition, 2000.

D: *Daybreak*. 1881. R.J. Hollingdale, trans., Maudemarie Clark and Brian Leiter, ed. New York: Cambridge University Press, 1997.

EH: *Ecce Homo*. 1888. In W. Kaufmann, trans. and ed., *The Basic Writings of Nietzsche*. New York: Modern Library Edition, 2000.

GM: *On the Genealogy of Morality*. 1887. Clark, Maudemarie, and Swenson, Alan J, trans. Indianapolis: Hackett Publishing Company, 1998.

GS: *The Gay Science*. 1882/1887. W. Kaufman, trans. New York: Vintage Books, 1974.

HA: *Human, All Too Human*. 1878. R.J. Hollingdale, trans. Cambridge: Cambridge University Press, 1996.

TI: *Twilight of the Idols*. 1888. In W. Kaufmann, trans. and ed., *The Portable Nietzsche*. New York: Viking Press, 1954

WP: *The Will to Power*. W. Kaufmann trans. New York: Random House, 1968.

WS: *The Wanderer and His Shadow*. 1878. R.J. Hollingdale, trans. Cambridge: Cambridge University Press, 1996.

Z: *Thus Spoke Zarathustra*. 1883-5. In W. Kaufmann, trans. and ed., *The Portable Nietzsche*. New York: Viking Press, 1954.

Other Works

Alfano, Mark. Forthcoming. *Nietzsche's Moral Psychology*. Cambridge: Cambridge University Press.

Anderson, R. Lanier. 2013. *Nietzsche on Autonomy*. In Gemes and Richardson, ed., *The Oxford Handbook on Nietzsche*. Oxford: Oxford University Press.

___. 2012. "What is a Nietzschean Self?" In Janaway and Robertson, ed., *Nietzsche, Naturalism, and Normativity*. Oxford: Oxford University Press.

Clark, Maudemarie. 2015. "Nietzsche on Free Will, Causality, and Responsibility." *Nietzsche on Ethics and Politics*, 75-96. Oxford: Oxford University Press.

___. 1998. "On Knowledge, Truth, and Value: Nietzsche's Debt to Schopenhauer and the Development of Empiricism." In Janaway ed., *Willing and Nothingness: Nietzsche as Schopenhauer's Educator*. Oxford: Clarendon Press, 37-78.

___. 1994. "Nietzsche's Misogyny." *Nietzsche on Ethics and Politics*, 141-50. Oxford: Oxford University Press.

Clark, Maudemarie and Dudrick, David. 2012. *The Soul of Nietzsche's Beyond Good and Evil*. Cambridge: Cambridge University Press.

Chisholm, Roderick. 1964. “Human Freedom and the Self.” In Watson ed., *Free Will*, New York: Oxford University Press, 26-37.

Frankfurt, Harry. 1987. "Identification and Wholeheartedness." *The Importance of What we Care About*, 159-176. Cambridge: Cambridge University Press, 1988.
__.1971. "Freedom of the will and the concept of a person." *The Importance of What we Care About*, 11-25. Cambridge: Cambridge University Press, 1988.

Gemes, Ken. 2006. “Nietzsche on Free Will, Autonomy, and the Sovereign Individual.” *Proceedings of the Aristotelian Society*, 80 (1): 321–38.

Janaway, Christopher. 2012. “Necessity, Responsibility, and Character: Schopenhauer on Freedom of the Will.” *Kantian Review* 17 (3): 431-57.
__. 2007. *Beyond Selflessness*. Oxford: Oxford University Press.
__. 2006. “Nietzsche on Free Will, Autonomy, and the Sovereign Individual.” *Proceedings of the Aristotelian Society*, 80 (1): 339–57.

Katsafanas, Paul. 2016. *The Nietzschean Self*. Oxford: Oxford University Press.

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

Leiter, Brian. 2015. *Nietzsche on Morality* (3rd Edition). London: Routledge.
__. 2007. “Nietzsche’s Theory of the Will.” *Philosopher’s Imprint* 7 (7): 1-15.
__.2002. *Nietzsche on Morality* (2nd Edition). London: Routledge.

McKenna, Michael and Pereboom, Derk. 2015. *Free Will: A Contemporary Introduction*. London: Routledge.

Riccardi, Mattia. 2015. “Inner Opacity: Nietzsche on Introspection and Agency.” *Inquiry* 58 (3): 221-43.

Richardson, John. 2004. *Nietzsche’s New Darwinism*. Oxford: Oxford University Press. __1996. *Nietzsche’s System*. Oxford: Oxford University Press.

Ridley. 2007. “Nietzsche on Art and Freedom.” *European Journal of Philosophy* 15 (2): 204-24.

Rutherford, Donald. 2011. “Freedom as a Philosophical Ideal: Nietzsche and His Antecedents.” *Inquiry* 54 (5): 512-40.

Stern. 2013. “Nietzsche, Amor Fati, and *The Gay Science*.” *Proceedings of the Aristotelian Society*, 113 (2.2): 145–162.

Schopenhauer. 1844. *The World as Will and Representation: Volume II*. E.F.J. Payne, trans. New York: Dover, 1958.
__. 1839. *On the Freedom of the Will*. Konstantin Kolenda, trans. Mineola: Dover, 2005.
__. 1818. *The World as Will and Representation: Volume I*. Norman, Welchman, and Janaway, trans. and ed., Cambridge: Cambridge University Press, 2010.

Snelson, Avery. 2017. “The History, Origin, and Meaning of Nietzsche’s Slave Revolt in Morality.” *Inquiry* 60: 1-30.

Taylor, Richard. 1962. *Philosophical Review* 71 (1):56-66.

Van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

Watson, Gary. 1975. “Free Agency.” *Journal of Philosophy* 72: 205-20.

Chapter 2

Nietzsche's Strawsonian Reversal

Abstract: The second essay of the *Genealogy* (GM II) is proclaimed by Nietzsche to be the "long history of the origins of *responsibility*," but the immediate context in which this claim is made, coupled with GM II's broader aims and themes, makes interpreting this claim immensely difficult. Specifically, Nietzsche endorses an ideal of responsibility in relation to the sovereign individual, whereas the rest of the essay is concerned with providing an account of the origins of conscience, bad conscience, and guilt—that is, an account of moral capacities that most humans possess. More problematically, this account appears to be inconsistent with Nietzsche's persistent and emphatic denial of free will and moral responsibility throughout his works. I first consider and critique two kinds of strategies employed in the literature for addressing this issue. I then propose a Strawsonian reading of GM II that I argue does a better job of solving it. Specifically, I argue that GM II analyzes responsibility naturalistically in terms of the *practice* of holding oneself and others responsible, as constituted by what Nietzsche calls the "reactive affects," primarily guilt. As for Strawson, a consequence of this is that an analysis of free will is not necessary to understand the nature and conditions of responsible agency.

Keywords: responsibility, free will, Strawson, exemption, bad conscience, guilt

I. The Interpretive Challenge of GM II

Nietzsche's skepticism surrounding free will and moral responsibility is prominent throughout his works. Many of these contexts suggest, implicitly or explicitly, that libertarian free will is a necessary condition for moral responsibility, and since we do not possess free will in this sense, we are not morally responsible. We find Nietzsche endorsing this line of reasoning throughout *Human, All too Human*, for instance. There he claims that moral responsibility "rests on the error of freedom of will," when in fact "man can be made accountable for nothing, not for his nature, nor his motives, nor his actions, nor for the effects he produces," because these are "altogether a necessary consequence of … things past and present" (HA I: 39).[48] Similarly, in *Daybreak* he suggests we are just as responsible for our actions as we are our dreams (D 128), and in

[48] Note, throughout HA Nietzsche relies on Schopenhauer's incompatibilist conception of free will as the "*liberum arbitrium indifferentiae*," the "free choice of indifference."

Beyond Good and Evil he claims libertarian or "causa sui" free will is "the best self-contradiction that has been conceived so far" (BGE 21). Later Nietzsche even speculates that libertarian free will and agent causation have their roots in the affect *ressentiment* (GM I: 13). Finally, continuing this theme, in *Twilight of the Idols* he tells us that "the doctrine of the will has been invented especially for the purpose of punishment, that is, because one wanted to impute guilt" (TI VI: 7).

This brief summary is by no means exhaustive of the contexts where Nietzsche expresses skepticism about our possessing free will or our being morally responsible. To the contrary, his works are replete with statements to this effect. I mention them here only to point out how jarring they are when viewed in the context of the second essay of the *Genealogy of Morality* (GM II). There Nietzsche proclaims to offer "the long history of the origins of *responsibility*" (GM II: 2), and unlike these other context, there he nowhere that we are responsible on the basis that we lack free will. In fact, not only does GM II neglect the topic of free will almost entirely, Nietzsche seems to offer a naturalistic explanation of why humans *are* responsible. He does this by offering a developmental account of the origins of conscience, bad conscience, and guilt—that is, an explanation of the emergence of psychological capacities plausibly believed to lie at the heart of moral agency. Also, quite apart from arguing that the will is merely epiphenomenal (*contra* Leiter [2007]), Nietzsche argues that the conscience is the will's "memory," and as such it enabled us to will diachronically. He describes this faculty as

> [A]n active no-longer-wanting-to-get-rid-of, a willing on and on of something one has once willed, a true *memory of the will:* so that a world of new strange things, circumstances, and even acts of the will may be placed without reservation between the original "I want," "I will do," and the actual discharge of the will, its *act*, without this long chain of the will breaking. (GM II: 1)

If this weren't enough, in the next aphorism Nietzsche proceeds to praise the enigmatic "sovereign individual" as a paragon of "freedom," "autonomy," and "responsibility" (GM II: 2).

How are we to make sense of these conflicting remarks? That is, how can we reconcile Nietzsche's "long history of the origins of *responsibility*" with his persistent and emphatic denial of free will and moral responsibility throughout his other works? Call this the *Interpretive Challenge* of GM II. My aim in this essay is to offer a solution to this problem. Before doing so, in Section 2 I will briefly consider and critique two types of *quarantine strategies* employed in the secondary literature for addressing it. I call them "quarantine" strategies because they attempt to render GM II's analysis of responsibility immune from the implication that human beings are, in general, *morally* responsible. Some scholars do this while trying to recuperate a positive notion of responsibility (or "freedom" or "autonomy") that Nietzsche attributes to the sovereign individual. While others offer deflationary interpretations of the sovereign individual and the ideas of "freedom," "autonomy," and "responsibility" attributed to him. Both strategies are problematic on independent grounds, and so I will propose an alternative reading of GM II's analysis of responsibility in Section 3.

Broadly, what I propose is an alternative *framework* for interpreting GM II. I will argue that its analysis of responsibility belongs in the lineage of sentimentalists like Hume and P.F. Strawson, who privileged the moral emotions to explain the nature and conditions of responsible *agency* in terms of our moral psychology.[49] Specifically, I will argue that Nietzsche, like Strawson in "Freedom and Resentment" (1962), "reverses" the

[49] Paul Russell (1995) argues for a "naturalistic interpretation" of Hume's compatibilism inspired by Strawson's analysis of responsibility, which privileges our moral psychology as opposed to an a priori conceptual analysis of the ideas of "liberty" and "necessity." I will argue for a similar reading of GM II here.

traditional or metaphysical interpretation of responsibility in GM II. This means that, rather than analyzing responsibility in terms of a logical or conceptual analysis of the relationship between free will and determinism, he analyzes it in terms of the *practice* of holding oneself and others responsible, as constituted by what Strawson calls the "reactive attitudes" and Nietzsche calls the "reactive affects" (GM II: 11). I will further suggest that this commonality is supported by their shared commitment to provide a *naturalistic* analysis of responsibility in their respective essays.

On my interpretation, GM II articulates a view of responsible *agency*, or moral agency, that does not depend on an analysis of free will. What distinguishes responsible or moral agents from other kinds of creatures is both a matter of moral motivation, a matter of being motivated to act for moral reasons or motives, and an ability to *hold oneself* responsible. I will argue that the development of bad conscience and the ability to feel guilt qualifies us as moral agents in this sense. However, our becoming moral agents also made us appropriate targets of the moral reactive attitudes, and hence blame, and so in that sense at least we *are* morally responsible. Finally, I offer my solution to the Interpretive Challenge in Section 4, where I relate this account of responsible agency to Nietzsche's critique of libertarian free will and responsibility. There I will suggest that, while he does unequivocally reject libertarian free will as a condition of responsibility, and though this coupling is essential to the kind of morality he wished to "overcome," the implications of this are not as far-reaching as they might seem.

II. Two Kinds of Quarantine Strategies

Nietzsche's claim that GM II is "the long history of the origins of *responsibility*" is difficult to interpret for several reasons, beginning with the fact that the surrounding context in which this claim is made offers divergent and conflicting answers concerning

its *scope*. In the following sentence Nietzsche equates this "history," and so its associated idea of "responsibility," with "the task of breeding an animal that is *permitted to promise*," who he then identifies as the "sovereign individual" (GM II: 2). The implication is that being "responsible" and having the status of a "permitted promisor" are synonymous, which Nietzsche further specifies is not simply a matter of being regular and predictable in one's behavior, or having developed stable disposition to keep one's promises. A person must instead "earn" the "trust" they invite (GM II: 2), or be *trustworthy*, and, as Nietzsche describes it, being worthy of the trust of others is no minor feat. It involves not only conveying the solemnity of one's promise, and an ability to will diachronically, but autonomy and a commitment to personal integrity. Thus, Nietzsche appears to regard trustworthiness as a kind of *ideal*, something perhaps that all humans are capable of, but few manage to live up to.[50]

Yet there is also a more familiar idea of "responsibility" pervading GM II, preliminary to and presupposed by this ideal of trustworthiness, and which would actually seem to be Nietzsche's primary focus in the text. The opening lines of GM II presumably reference this notion of "responsibility," because it begins by asking a question concerning permitted promisors whose scope ranges over *all* human beings: "To breed an animal that is *permitted to promise*—isn't this precisely the paradoxical task nature has set for itself with regard to man? Isn't this the true problem *of* man? …" (GM II: 1). There can be no doubt that this "problem" or "task" applies to *all* human beings—to "man" *in general* as opposed to other animals. Nietzsche even claims that this problem or task "has been solved to a high degree" (GM II: 1). From this we ought to

[50] See Chapter 5 for analysis. Several remarks Nietzsche makes in the second aphorism point in this direction. Space precludes me from analyzing them here, but for interpretations of this kind, see Anderson (2013), Rutherford (2011), and Dannenberg (2017).

infer that most humans are "permitted promisors," perhaps in a lesser sense than sovereign individuals, but at least in the sense that they are not merely regular and predictable, but indeed "responsible."[51]

These conflicting remarks present a special problem for what is by far the most common variant of quarantine strategy. Generally, proponents of this strategy try to recuperate a positive notion of responsibility in connection to the sovereign individual—or an account of "agency,"[52] or an *ideal* of freedom or autonomy,[53] or an account of trust ("permitted promising")[54]—but they maintain either that these accounts are *unrelated* to moral responsibility, or that they are *ideals*, and as such do not generalize to all human beings. (Often a scholar's interpretation will be committed to both of these claims.) In pursing either strategy many of these scholars do not address the claim that GM II provides a "history of the origins of *responsibility*," and sometimes do not even address the claim that the sovereign individual is "responsible." Fortunately, Ken Gemes is an exception in the latter respect. According to him:

> Nietzsche rejects deserts free will and affirms agency free will. Nietzsche wants to reject the notion that in doing such-and-such one might have done otherwise, yet he wants to affirm that genuine agency is possible, if only for a select few. It is this that explains why in some contexts he denies free will and in others positively invokes free will. The denials are denials of deserts free will and the invocations are invocations of agency free will (Gemes 2006: 322).

Gemes helpfully goes on to explain that there are two different notions of responsibility tied to each idea of free will: "To say that so-and-so is responsible for such-and-such can mean that they deserve punishment/reward for it. On the other hand to say that

---

[51] Several remarks that Nietzsche makes in second and third aphorisms point in this direction. Again, space precludes analysis here, but for interpretations of this kind see Reginster (2011, 2017) and Zamosc (2011).

[52] See Gemes (2006), Migotti (2013), and Poellner (2009).

[53] See Anderson (2013), Rutherford (2011), also Gemes (2006).

[54] See Dannenberg (2017), Reginster (2011, 2017), also Migotti (2013).

someone is responsible for such-and-such can simply mean that it was their doing" (Gemes 2006: 322).

Read in these terms, GM II is concerned with something like what Gary Watson (1996) called the "aretaic" face of responsibility, or responsibility as attributability, as opposed to responsibility as accountability. Roughly, someone is responsible in the sense that the action is *attributable* to her if it is expressive of who she is, or what she stands for on matters of value. One is *accountable*, on the other hand, if sanctions are applicable and fairly applied as a result of the agent's action. The latter notion of responsibility is inseparable from considerations of desert because it is concerned with just punishment and reward, whereas the former is presumably separable from considerations of desert entirely. Gemes is certainly right that GM II is concerned with the issue of attributability. Indeed, the idea of "autonomy" that Nietzsche attaches to the sovereign individual is naturally interpreted in such terms. He possesses "a long unbreakable will, [and] has in this possession his *standard of value* as well" (GM II: 2). If the sovereign individual takes a stand on matters of value and acts accordingly, then such actions are indeed expressive of who he takes himself to be.

Yet the distinction between attributability and accountability is problematic when applied to GM II as whole. The reason for this is that GM II is *overwhelmingly* concerned with sanctions and holding others (and eventually oneself) to expectations of various kinds. Initially, in the morality of custom, humans were held to expectations to conform their behavior to customary rule prohibitions ("I will nots" [GM II: 3]), and punished when they failed to do so. This is Nietzsche's account of the origins of

conscience, which developed as a "memory" these rules.[55] Later, in creditor-debtor relationships, punishment did not merely serve the function of establishing and enforcing norms; it was administered in accordance with the principle of "equivalence" (GM II: 4), the idea that offenders deserve to suffer in proportion to the harm they caused. Nietzsche claims that equivalence was "the oldest and most naïve canon of moral *justice*" (GM II: 8), because it operated according to primitive and evolving ideas of fairness and desert.[56] Finally, bad conscience did not originate as a consequence of punishment, but its development meant that we began punishing *ourselves*. So not only is the accountability face of responsibility central to understanding GM II's "long history" of responsibility, it is unclear how responsibility as attributability might emerge within this story *apart* from it.

Proponents of this kind of quarantine strategy therefore face a significant interpretive burden. Not only must it be shown that Nietzsche's positive idea of responsibility in connection to the sovereign individual is about A (responsibility as attributability, genuine agency, an ideal of freedom and/or autonomy, or trust), and how that does not implicate B (responsibility as accountability), it must *also* be shown how A arises within GM II's "long history" of responsibility, which is overwhelmingly concerned with B. The sovereign individual does not stand outside of this "history." He is not the product of divine intervention or immaculate conception, and so even if he represents an *ideal* of responsibility—because he possess the "highest form" of conscience (GM II: 3), or because conscience is his "dominant instinct" (GM II: 2)—it's still the case that he has been subject to the same forces that shaped us lesser mortals. Put another way, since

[55] See Snelson (2019) and Chapter 3 for analysis.

[56] As Maudemarie Clark has argued, retributive punishment emerges in Nietzsche's story along with equivalence, at which point it becomes a "primitive form of moral address" (2015b: 93).

this ideal of responsibility has been ineluctably informed by the practices that contributed to the development of conscience, by the "long history," the proponent of this strategy must also provide a plausible explanation of the sovereign individual's place within that history, in particular GM II's genealogy of guilt. However, most interpreters do not do this, and so the distinction they draw between A and B ends up being more *stipulative* than substantive.[57]

The second kind of quarantine strategy I will now consider maintains that GM II's analysis of conscience, though it is directly related to the development of moral capacities that all humans possess, is insufficient to justify ascriptions of *genuine* responsibility. In other words, Nietzsche means "responsible" only in a revisionary or highly attenuated sense. I call them "deflationary" quarantine strategies for this reason. Christa Davis Acampora (2006), Lawrence Hatab (2009), and Brian Leiter (2011) all offer quarantine strategies of this kind. Here is a statement of Leiter's position, which he makes in reference to the sovereign individual:

> And what exactly is the 'trick' of this well-trained animal? Surely it bears emphasizing that he is described as having one and only one skill: he can actually make and keep a promise! And why can he do that? Because he can remember that he made it, and his behavior is sufficiently regular and predictable, that others will actually act based on his promises. … And the 'conscience' of this self-important creature, as Nietzsche makes clear, consists in nothing more than the ability to remember his debts. (2011: 108)

These authors are unified in the belief that the formation of conscience is insufficient to justify blame and ascriptions of responsibility. According to Leiter, the conscience "consists in nothing more than an ability to remember [one's] debts" (2011: 108).

[57] Reginster (2017) is a notable exception.

Therefore, the sovereign individual is "permitted to promise" because he is "steady enough to make a promise and honor it" (Leiter 2011: 108).[58]

The main problem with these deflationary readings is easy to identify. Succinctly, they under-describe the capacities conferred by the conscience *qua* "memory of the will," and they therefore conflate the distinction between mere reliability and responsibility that Nietzsche himself invokes. Consider, first, a context where Nietzsche identifies reliability of behavior as a "presupposition" of the "memory of the will":

> But how much this [memory of the will] presupposes! In order to have this command over the future in advance, man must first have learned to separate the necessary from the accidental occurrence, to think causally, and anticipate what is distant as if it were present, to fix with certainty what is end, what is means thereto, in general to be able to reckon, to calculate,—for this, man himself must first of all have become *calculable, regular, necessary*, in his own image of himself as well, in order to vouch form himself *as future*, as one who promises does! (GM II: 1)

Of course, if reliability is a *presupposition* of the memory of the will, then reliability is not the capacity conferred *by* the memory of the will.[59] To understand why this is, we must consider Nietzsche's genealogy of conscience in a bit more detail.

Conscience goes through various transformations and permutations in GM II. It has behind it "a long history and metamorphosis" (GM II: 3), as Nietzsche says. One of these forms of conscience, indeed the original form described by him, is nothing more than a "memory" of the social expectations of others, of the rules they imposed on us. As noted, this form of conscience originated in the morality of custom as a memory of "I will

---

[58] Similarly, Acampora believes Nietzsche wants to explain a *Kraft*; specifically, how we became "capable of" making promises by remembering them (2006: 149). Hatab (2009) endorses her analysis.

[59] While Leiter (2011: 106) acknowledges this, because he also believes that the conscience is only an ability to remember one's promises, it confers nothing more than the same behavioral capacity, i.e., an ability to rely on others to do what is expected of them. See Chapter 5, §2.

nots" (GM II: 3), and was inculcated in us through punishment. This did indeed make us reliable, or regular and predictable in our behavior, because the formation of conscience allowed us to remain *aware* of the rules others imposed on us and enforced through punishment. But Nietzsche is clear that being a reliable agent is not the same thing as being responsible:

> The enormous work of what I have called the "morality of custom" ... the true work of man on himself for the longest part of the duration of the human race, his entire *prehistoric* work, has in this its meaning, its great justification ...with the help of the morality of custom and the social straightjacket, man was *made* truly calculable. If, on the other hand, we place ourselves at the end of the enormous process ... where society and its morality of custom finally brings to light that *to which* it was only the means: then we will find as the ripest fruit on is true the *sovereign individual* ... (GM II: 2)

Again, if the morality of custom was only a "means" to the sovereign individual and it made us “calculable,” then reliability and responsibility cannot amount to the same thing.

To explain this, let's consider Nietzsche's description of the "memory of the will." This type of "memory," he says, is "an active no-longer-wanting-to-get-rid-of," and "thus by no means simply a passive no-longer-being-able-to-get-rid-of" (GM II: 1). Imagine that I make a promise to you. The former kind of "memory" would imply that I keep this promise because I in some sense *want* to keep it, perhaps because I "identify" with it, or because I *value* keeping my promises, or because doing so is important to my integrity. The second kind of "memory," on the other hand, would imply that I keep my promise simply because I cannot rid myself of thought of what might happen if I don't. Perhaps I am worried about my reputation, or I fixate on how breaking my promise would hurt your feelings, or, as in the case of creditor-debtor relationships, I am scared of the cruelty you will inflict on me in recompense.

As these examples illustrate, the difference between these two types of "memory" consists in the *motivation* associated with each. In the first case the motive is in *internal* to the agent and self-generating, whereas in the second case it remains *external*. This distinction between internal and external sources of motivation might be understood in two ways. We might say that external motives are essentially concerned with the *consequences* of keeping or failing to honor one's obligations, whereas internal motives are not.[60] Or we might say that, when internally motivated, the agent would still fulfill her obligation in the absence of incentives, whereas an agent who is only externally motivated would not. In either case, the kind of motivation associated with this "highest form" (GM II: 3) of conscience is supposed to be internal to the person and self-generating. It is not essentially connected to the expectations of others.

So, we can conclude that these deflationary readings are inadequate on textual grounds. They attempt to show that when Nietzsche says "responsible," he means this only in a deflationary or attenuated sense, equivalent to being reliable, or "regular" and "predictable" (GM II: 2) in the context of promise-keeping. However, because Nietzsche himself distinguishes between being reliable and being responsible, we will need to find an alternative solution to the Interpretive Challenge of GM II.[61]

III. Nietzsche's Strawsonian Reversal

It's no secret that the philosophical topic of responsibility has traditionally been analyzed in terms of free will and its relationship to determinism. Following R. Jay Wallace, I will

[60] See BGE 32's remarks on the morality of custom: "During the longest part of the human history—so-called prehistoric times—the value or disvalue of an action was derived from its consequences."

[61] Note, even a "persuasive definition" (Leiter 2011: 112) of responsibility must still render coherent the distinction between mere reliability and responsibility, and the forms of conscience that apply to each, or else Nietzsche was simply mistaken to draw these distinctions in the first place.

call this the *metaphysical interpretation* of responsibility. "On this interpretation," he explains, "we would suppose that there is a fact of the matter about responsibility 'in itself,' a fact about what it is to be *genuinely* or *really* responsible, and that this fact is prior to and independent of our practice of treating people as morally responsible agents" (Wallace 1994: 87). According to this approach, being "responsible" is primarily a matter of a priori conceptual analysis and depends on the idea of free will, which signifies a relationship of control between the agent and her action such that she can be legitimately praised or blamed for it. "Control" is defined in terms of whether the agent is able to exercise causal power over a range of alternatives, resulting in a "picture of responsibility as consisting in a kind of freedom of choice" (Wallace 1994: 86).

There is abundant evidence that Nietzsche was *not* thinking of responsibility in these terms in GM II. Note, first, that GM II's idea of responsibility has a "long history," i.e., a developmental record that can be unearthed through genealogical investigation. Such an approach could not be more antithetical to the metaphysical interpretation, which instead concerns where an agent has freedom of choice in a particular case. More importantly, GM II shows practically no concern for the relationship between free will and determinism. There are no references to determinism therein (though there is a passing reference to "unfreedom of the will" [GM II: 21]), and only three references to free will. Nietzsche attributes free will to the sovereign individual (GM II: 2). We will consider shortly and at length another passage where Nietzsche claims that libertarian free will is dispensable to the practice of holding others responsible (GM II: 4). Finally, the third reference speculates that philosophers invented the concept of free will (GM II: 7). In short, as Bernard Reginster has observed, "'freedom of the will' plays no significant role" in GM II (2011: 69).

On the other hand, the development of conscience bears a direct and obvious connection to "responsibility," when one instead understands that in terms of the *practice* of holding others to expectations, along with the proneness to experience emotional reactions when they go unmet, coupled with the development of uniquely moral capacities (i.e., bad conscience and guilt). That is precisely how Nietzsche approaches the topic of "responsibility" in GM II. Before analyzing GM II's approach further, I wish to briefly explain why I believe Nietzsche was led in this direction.

Quite apart from the relationship between free will and responsibility, Nietzsche had long been interested in the phenomenon of *moral agency*. Being a moral agent is a matter of being motivated to act in a certain way, a matter of being motivated to act for moral reasons or motives, but this kind of motivation also makes moral agents "responsible" in a way that non-moral agents are not. Specifically, they are "responsible" in the sense that they are *responsive* to moral reasons, a receptivity that is grounded in their ability to experience moral emotions like guilt, and so they can conduct themselves in ways that non-moral agents cannot. It should be clear from GM II's developmental account of conscience, bad conscience, and guilt that it is very much preoccupied with moral agency understood in these terms. However, just as noteworthy is the fact that GM II also represents a significant *shift* in Nietzsche's thinking about moral agency compared to his previous works.

Following Kant and Schopenhauer, Nietzsche had long maintained that to act morally one must act from a moral *motive*. In *Human, All too Human* he argued that we were incapable of acting from such motives, presumed to be purely “unegoistic,” and

therefore moral agency was impossible.[62] In *Daybreak*, however, he reversed course. He now believed that acting from moral motives was possible (D 103), but that doing so did not make us responsible. The reason being that acting from this motive of fearful "obedience to tradition" (D 9) was little more than an instinct or habit inculcated in us during the morality of custom. In each of these works Nietzsche assigned a prominent role to the morality of custom to explain moral agency, but by the time we get to the *Genealogy* it plays only a subsidiary role. Customs are now understood to be non-moral rules, whereas acting in accordance with *moral* rules implicates bad conscience and guilt.[63] As we saw above, the morality of custom now plays only an instrumental role in explaining what I have called "reliable agency," which is a precursor to responsible agency (being "permitted to promise").

It is my contention that this shift in Nietzsche's thinking about moral agency explains his shift away from the metaphysical interpretation of responsibility in GM II. I propose to defend this supposition by drawing on a crucial commonality between P.F. Strawson's approach to responsibility in "Freedom and Resentment" and Nietzsche's in GM II. In their respective essays, both Strawson and Nietzsche provide a *naturalistic* analysis of responsibility, thereby "reversing" the order of explanation assumed by the metaphysical interpretation, as this pertains to the justification of the reactive attitudes.

3.1 Strawson's Naturalism and the "Reversal" Thesis

What does it mean to say that an approach to responsibility is "naturalistic?" Most interpreters of Strawson have said something to this effect, and they have done so

[62] See especially HA I: 34, 96-99.

[63] As Clark observes, "Rules obeyed only out of fear or instinct are not yet perceived as moral rules by those who are disposed to obey them. The main question Nietzsche pursues in *GM* II concerns how such non-moral rules, laws, and customs were transformed into moral ones. His basic answer is that this happened through the development of guilt" (2015a: 68).

moreover by drawing a contrast to the metaphysical interpretation. Consider these remarks from Paul Russell:

> Strawson's strategy is to take what may be described as a "naturalistic turn." Rather than asking directly, in the abstract, what *is* a responsible agent, Strawson suggests that we should consider in more detail, with more precision, what is involved in the attitudes that we take toward those who we regard as responsible agents. That is to say, what is involved in *holding* a person responsible? An approach of this kind depends less on a conceptual analysis of "freedom" and more on a descriptive psychology of human moral emotions. (2011: 200)

This commitment to naturalism is exemplified by Strawson and Nietzsche because they both analyze responsibility as a *social phenomenon*, as a set of practices that is grounded in our expectations of others and the emotions elicited in us when those expectations go unmet. Strawson calls these emotions the "reactive attitudes," whereas Nietzsche calls them "reactive affects" (GM II: 11). According to Strawson, they "rest on, and reflect, an expectation of, and demand for, the manifestation of a certain degree of goodwill or regard on the part of other human beings toward ourselves" (1962: 84), and as such they constitute our practice of blame, of holding ourselves and others responsible. Though Nietzsche's "reactive affects" are not coextensive with Strawson's "reactive attitudes," they fit this same general pattern.[64]

Importantly, Strawson also draws our attention to two kinds of considerations or "pleas" that typically cause us to mollify or suspend these attitudes, and thus mollify or withdraw entirely attributions of responsibility. These have come to be known as

[64] In his most extensive discussion of the "reactive affects," Nietzsche contrasts them with the "*active* affects like desire to rule, greed, and the like" (GM II: 11). Reactive affects, by contrast, are emotional *reactions to* others or states of affairs in which an agent suffers some "injury," emotions like anger, revenge, and *ressentiment*. Guilt is a reactive affect, on my reading, because it is a response to the agent inflicting this injury upon *himself*. Reactive affects are therefore responses to holding others (and eventually oneself) to expectations, but these need not be specifically *moral* expectations, as they are for Strawson, as I elaborate below.

"excuses" and "exemptions,"[65] each corresponding (roughly) to the distinction between being responsible for an action and being a responsible agent in general.[66] Understood in these terms, when one is *excused* from blame she is still seen as a responsible agent, or one whom the expectation or demand for good will or regard is maintained, but the attitudes that typically accompany the violation of this demand, like resentment, are mollified or temporarily suspended. That is, we withdraw blame for the *action*, because we discover that it did not possess the ill or indifferent will it was initially thought to possess, while still regarding the person as "a term of moral relationships" (Strawson 1962: 86). *Exemptions* differ from excuses in that they involve the withdrawal of this demand or expectation itself, and so involve treating the person as one who is *not* a responsible agent. Strawson believes that in these cases we instead adopt the "objective attitude," by treating the agent as "an object of social policy," as "a subject for treatment," to be "managed or handled or cured or trained; perhaps simply to be avoided" (1962: 79). In other words, we regard the agent in a similar manner as we would treat toddlers or pets.

By making the reactive attitudes central to our understanding of responsibility, many have suggested that Strawson "reverses" the metaphysical interpretation of responsibility. As Gary Watson initially articulated this idea:

> In Strawson's view, there is no such independent notion of responsibility that explains the propriety of the reactive attitudes. The explanatory priority is the other way around: It is not that we hold people responsible because they *are* responsible; rather, the idea (*our* idea) that we are responsible is to be understood by the practice, which itself is not a matter of holding some

[65] "Excuse" and "exemption" are not Strawson's terms, but Gary Watson's (1987). However, they are helpful labels for what Strawson calls "type-1" and "type-2" pleas (see Strawson 1962: 77-8).

[66] See especially Strawson's discussion of the "vicarious analogues" of the first-personal reactive attitudes (1962: 85).

> propositions to be true, but of expressing our concerns and demands about our treatment of one another. (1987: 222)

According to Watson, this "reversal" thesis is concerned with the justification or "propriety" of the reactive attitudes. Whereas the metaphysical interpretation assumes this justification is *independent of* the practice of holding responsible, assuming we have a prior conception of what it means to *be* responsible (through the ideas of free will and determinism), Strawson contends, to the contrary, that we must seek this justification *within* that practice, as constituted by the reactive attitudes and the two types of "pleas" just considered. Succinctly put, Strawson's *methodological commitment* to naturalism establishes a relationship of *conceptual dependence*, the main consequence of which is that an analysis of free will is not necessary to understand the nature and conditions of responsible agency.[67]

Strawson's clearest expression of this commitment comes near the end his essay. "Only by attending to this range of attitudes," he claims, "can we recover from the facts as we know them a sense of what we mean, i.e. of *all* we mean, when speaking the language of morals, we speak of desert, responsibility, guilt, condemnation, and justice" (Strawson 1962: 91). One cannot help but notice that the idea of free will is conspicuously absent. As Michael McKenna and Derk Pereboom note, "The force of Strawson's essay … *cannot* be understood in terms of a specific line of argumentation for or against any single proposition regarding free will or moral responsibility. Moreover, it is not altogether clear or easy to state how it is that Strawson's overall position bears on any other single point regarding the freedom of the will" (2016: 142). One is tempted to

[67] Hieronymi (forthcoming) calls this the "broadly Wittgensteinian interpretation" (1) of Strawson's argument.

think this is because he, like Nietzsche, thought it dispensable to analyzing responsible agency.

Regrettably, I cannot comment further on Strawson's argument or its success here. Suffice it to say, while all commentators of Strawson agree that his approach is "naturalistic" on account of its privileging the reactive attitudes, not all interpret him as advancing this "reversal" thesis,[68] and some have argued that this thesis does not succeed in making his argument anti-libertarian.[69] We must also stress the fact that Nietzsche's aim in GM II was not to "reconcile" the debate between incompatibilists and compatibilists, both of whom according to Strawson "overintellectualize the facts" as they relate to responsibility (1962: 91). Yet Nietzsche undoubtedly did share Strawson's *methodological commitment* to naturalism in GM II, and we shall see momentarily that this established a similar relationship of *conceptual dependence*, specifically one that did not take free will to be a condition of responsible agency.

3.2 Nietzsche's "Reversal" Thesis

Nietzsche offers his version of the "reversal" thesis in the fourth aphorism of GM II. This aphorism is moreover critical to understanding GM II's analysis of responsibility. There he explains the origin of punishment "as *retribution*," which he maintains is the true and original form of the practice that "previous genealogists of morality" have failed to recognize. I will here assume that retributive punishment is defined by two features. First, it is a non-consequentialist, backward-looking response to wrongdoing, and secondly, it involves the belief that the offender "deserves" to suffer, and is therefore permissible (if not good) that he does.

---

[68] See Hieronymi (forthcoming), Wallace (1994), and Russell (2011).
[69] See Todd (2016).

As we have seen, not all the forms of punishment Nietzsche discusses in GM II satisfy these criteria. Punishment that inculcated the "I will nots" (GM II: 3) was not retributive, since its sole function was to make us reliable agents by imprinting a memory of the group's rules. Retributive punishment instead arose within creditor-debtor relationships, when creditors would punish debtors as a form of compensation, which required establishing an *equivalence* price between the initial injury and subsequent punishment. What those "previous genealogists" failed to recognize is that this practice "developed completely apart from any presupposition concerning freedom or lack of freedom of the will" (GM II: 4). Nietzsche explains:

> The thought, now so cheap and apparently so natural, so unavoidable, a thought that has even had to serve as an explanation of how the feeling of justice came into being at all on earth—"the criminal has earned his punishment *because* he could have acted otherwise"—is in fact a sophisticated form of human judging and inferring that was attained extremely late; whoever shifts it to the beginnings lays a hand on the psychology of older humanity in a particularly crude manner. Throughout the greatest part of human history punishment was definitely *not* imposed *because* one held the evildoer responsible for his deed, that is, *not* under the presupposition that only the guilty one is to be punished:—rather, as parents even today punish their children, from anger over an injury suffered, which is vented on the agent of the injury—anger held within bounds, however, and modified through the idea that every injury has its *equivalent* in something and can really be paid off, even if only through the *pain* of its agent. (GM II: 4)

How does this passage amount to a "reversal" of the metaphysical interpretation? First, note that Nietzsche believes the practice of retributive punishment is perfectly intelligible without taking into consideration the belief that one must have the ability to do otherwise. He proposes instead that we understand retributive punishment in terms of the *practice* of holding others to expectations, along with the proneness to experience emotional reactions when they go unmet. This is what Nietzsche's remarks about anger at the end of the passage are meant to signify. Anger in this context is a form of other-directed blame, an emotion one experiences in reaction to suffering some "injury," taking

as its object the perpetrator of the injury, expressed as a desire to harm him for the wrong committed. This involves holding the agent to affectively charged expectations. However, it is important to see that, as Nietzsche's remarks about children suggest, this form of the practice does *not* presume the offender is a responsible *agent*.[70]

According to Strawson, children are "creatures who are potentially and increasingly capable of both holding, and being objects of, the full range of human and moral attitudes, but are not yet truly capable of either" (1962: 88). This fact is reflected in our expectations and attitudes towards them. Parents hold their children to all kinds of expectations during their formative years—to not throw their food, to not climb the stairs, to pick up their toys, to be obedient in general, etc.—and violating any one of these expectations often does result in "punishment" or some kind of sanction. However, children are also among those whom we adopt the objective attitude most frequently. This does not mean we have no emotional reaction to their behavior, however, and Strawson similarly did not believe the objective attitude to be one of complete indifference. "It may be emotionally toned in many ways," he notes, "but not in all ways … it cannot include the range of reactive feelings and attitudes which belong to involvement or participation in inter-personal relationships" (1962: 79). So a mother may feel angry at, frustrated with, or disappointed in her child, if, say, he throws his

[70] GM II in fact describes four different forms of the practice of holding others and oneself to affectively charged expectations, each corresponding to different stages in the development of conscience. Holding others to *expectations of conformity*, an expectation to conform one's behavior to rules, the violation of which elicits anger (see GM II: 3, 9); holding others to *expectations of redress*, expectations of repayment connected to the notion of "equivalence," the violation of which elicits anger and eventually resentment as a response to unfairness (see GM II: 5, 10); holding *oneself* and others to *moral expectations*; expectations that offenders *ought* to *feel* guilt for violating personal obligations (see GM II: 8, 14-15); and finally, holding others to expectations to fulfill their promises from the participant stance. The first form of the practice is not moral; the second is a primitive form of moral address (because it involves considerations of desert, fairness, and justice); the third is moral blame; the fourth is trust.

dinner all over the floor, though she presumably will not feel resentment, indignation, or a sense of betrayal.[71] And the reason for this is that anger is elicited because the child disregards the parent's *authority*, not because he has in her view done anything *morally* wrong. The attitude, like the expectation, is non-moral.

We can now begin to make sense of Nietzsche's claim that the "evildoer," like the child, is punished but not held "responsible for his deed." He clarifies this by claiming that the evildoer is punished but "*not* under the presupposition that only the guilty one is to be punished" (GM II: 4). It is not immediately clear what "guilt" is meant to signify in this context, because "guilt" is ambiguous between two possible meanings. "Guilt" could stand for *objective guilt*, the fact of having committed an offense for which one is "guilty" and deserving of punishment (as in rendering a verdict of "guilty" in a court of law). Or the reference to "guilt" could stand for *subjective guilt*, the state a person is in when she *experiences* guilt, which is commonly attended by feelings of remorse, contrition, and diminished worth. Guilt in this latter sense depends on the existence of bad conscience because guilt is "that reaction of the soul called 'bad conscience, the 'pang of conscience'" (GM II: 14).

Not coincidentally GM II analyzes both kinds of guilt, but the primary notion Nietzsche is interested in explaining is subjective guilt. This is also the idea of guilt he opens the passage by discussing: "But how did that other 'gloomy thing,' the consciousness of guilt, the entire 'bad conscience' come into the world?" (GM II: 4). This question signals an important transition in GM II's analysis of conscience, from the non-

[71] Strawson thinks anger is a reactive attitude and as such incompatible with the objective attitude. However, Nietzsche's "reactive affects" form a broader class than Strawson's reactive attitudes, which is why I speak of anger in these terms. Relatedly, Hieronymi (forthcoming: 5) draws a helpful distinction between being "angry with" (someone), as opposed to "angry at" (something), and she takes the latter to be consistent with the objective attitude.

bad conscience as a memory of rules, to the bad conscience as the cause of guilt feelings. Asking this question leads Nietzsche to an analysis of creditor-debtor relationships because, in addition to bad conscience, subjective guilt has its roots in the idea of debt. Specifically, guilt is a feeling of self-punishment that aims at redress or repayment, but it is also for this reason just the internalized form of the "material" concept of debt (GM II: 4).[72] Roughly, objective guilt, which is just the agent's recognition of her *indebtedness to another*, is a condition of subjective guilt, which additionally involves *punishing oneself* for being in this state of indebtedness. The common thread between the two notions is punishment as a means of expiating one's guilt—of repaying one's debt—but subsequent to the formation of bad conscience the agent takes on the duty of repayment herself.

So, Nietzsche's claim in this passage is that retributive punishment emerged along with considerations of *indebtedness*, or objective guilt, as mediated by the principle of equivalence (the idea that offenders deserve to suffer in proportion to the harm caused). So, when he claims that punishment was administered but "*not* under the presupposition that only the guilty one is to be punished," he means that they were *objectively* guilty, but that this nonetheless does not qualify as an instance of holding "the evil-doer responsible for his deed." Why?

One possible answer is that he was not punished under the presupposition that he could have done otherwise, but not only does Nietzsche claim that contra-causal free will is dispensable to the practice of retributive punishment, he nowhere in GM II identifies free will as a condition of guilt *at all*. Naturally, this is because his focus is on subjective guilt, and its origins lie elsewhere, in the idea of debt and the development of

[72] Regrettably, I cannot expand on subjective guilt's relationship to the "moralization" of *Schuld* (GM II: 20), but I agree with Janaway (2007) and Reginster (2011) that moralization is the perversion of subjective guilt, not the process by which material debt *becomes* subjective guilt. See Chapter 4, §4-5 for analysis.

bad conscience, *not* considerations surrounding free will and determinism. The reference to "guilt" in this passage is therefore plausibly interpreted as a reference to *subjective guilt* (the idea of guilt it begins by exploring). If this is right, the sense in which the evil-doer is not held "responsible" is that he is not punished under the presupposition that he should *feel* guilty, because he does not have a bad conscience, and so is not liable to guilt, "the feeling of personal obligation" (GM II: 8). For these reasons he is not regarded as a responsible *agent*.

What makes the experience of guilt "personal" is also what makes the agent responsible or accountable for her actions. Later Nietzsche explains that

> The "bad conscience," this most uncanny and interesting plant of our earthly vegetation, did *not* grow … in the consciousness of the one's judging, the ones punishing, there was for the longest time *nothing* expressed that suggested one was dealing with a "guilty one." But rather with an instigator of injury, with an irresponsible piece of fate" (GM II: 14).

Likewise, the criminal "had no other 'inner pain'" (GM II: 14). "For thousands of years instigators of evil overtaken by punishment have felt *no different than Spinoza* with regard to their 'transgression': 'something has gone unexpectedly wrong here,' *not*: 'I should not have done that'—they submitted themselves to punishment as one submits to a sickness or a misfortune or death" (GM II: 15). In these contexts a "guilty one" is naturally understood to be a person who, by undergoing an episode of bad conscience, *holds himself* responsible or accountable. Similarly, in GM III Nietzsche tells us that a "*guilty* perpetrator" is "one who is receptive to suffering" (GM III: 15), meaning he takes punishment to be *deserved*.[73] Thus, the picture of responsible agency that emerges is one that is ultimately grounded in our peculiar receptivity to suffering, or, more

[73] Any animal with nociceptors is "receptive to suffering," but in that case it does not matter whether one is "guilty" or not. What it means to be "receptive to suffering" in reference to guilt is that one believes punishment to be an appropriate response to what one has done.

accurately, our willingness to inflict suffering on ourselves when we think it is merited or deserved, without which moral agency would be impossible.

Consider in this connection the above claim "I should not have done that" (GM II: 15). "Should" in this context is the moral-categorical should, not the prudential-hypothetical should. "If there was a critique of the deed back then," Nietzsche explains, "it was prudence that exercised this critique" (GM II: 15). In other words, the criminal's response was something like "I should not have gotten caught," or "I should be more careful next time," similar to how we might think "I should have gotten my flu shot." His thought was *not* "I should not have done that because it was (morally) wrong." The reason for this is that an episode of bad conscience is necessary to make the *moral* form of the ought-judgment possible, because the mechanism underlying bad conscience—internalization, the process in which our aggressive drives become directed toward the self (GM II: 16)—is necessary to produce feelings of remorse, contrition, and anger at oneself for doing what one did or being who one is.

Agents who are capable of feeling guilt are thus "responsible" for their conduct in a way that merely reliable agents are not. Not only can they modify their behavior in accordance with the expectations of others, by recognizing that transgressing established conventions or rules will result in their being punished. When agents who experience guilt violate these, they punish *themselves*, which also means that they *hold themselves* to *moral* expectations to honor them. And because they hold themselves to these moral expectations, they may also be legitimately held responsible by *others*. As we have seen, this requires holding others to expectations in a deeper and more significant sense than the readiness to punish them. Specifically, holding another responsible requires doing so under the presupposition that the offender is capable of and ought to *feel* guilty, or

ought to feel that she deserves punishment. Accordingly, when we hold others *morally* responsible, we do so not with the aim of deterrence, repayment, treatment, or training—all of which are compatible with the objective attitude. We do so with the aim of eliciting guilt, and we thereby regard the agent *as* one who is responsible.

IV. The Strawsonian Reply to the Interpretive Challenge

To restate the Interpretive Challenge of GM II we began with: How can we reconcile Nietzsche's "long history of the origins of *responsibility*" with his persistent and emphatic denial of free will and moral responsibility throughout his other works? If the preceding analysis has been right, there is no incompatibility between GM II and these other contexts because GM II's account of guilt and responsible agency does not depend for its legitimacy on an analysis of free will and determinism. Why is that? Consistent with our Strawsonian interpretation, our answer to this question must consider the two different dimensions along which one might be taken to be "responsible."

First, it is irrelevant, from the perspective of our moral psychology, whether one can be responsible if one does not possess free will, because whether one is a responsible *agent* is a matter of *competence* or moral capacity. That is, as far as the issue of *exemption* is concerned, whether one is deserving of blame hinges on whether she is capable of feeling guilt, on whether she is capable of *holding herself* responsible, on whether she is a moral agent, not whether she could have acted otherwise than she in fact did. Ok, but what about the issue of *excuses*, i.e., the issue of whether someone deserves blame for a particular action (presuming she is a responsible *agent*)? Here matters are considerably more complicated and made more complex by the shifts and development in Nietzsche's thinking about free will and responsibility in his late works.

Note, Nietzsche's remarks in the fourth aphorism suggest that excusing conditions track whether the action was "intentional," "negligent," or "accidental" (GM II: 4), none of which need to be analyzed in terms of free will. But, more importantly, there is plenty of evidence in the late works that should make us doubt that Nietzsche believed libertarian free will was a condition of responsibility. Of particular importance here is the fact that, as noted in Section 1, Nietzsche offers an account of the origins of libertarian free will and agent causation in GM I. There he argues that belief in libertarian free will was necessary to legitimize the slaves' blame of the nobles, to "affirm" themselves, and to express their *ressentiment* (GM I: 13). In *Twilight* he then tells us that this "priestly" conception of free will was invented "because one wanted to impute guilt" (TI VI: 7). Note, however, that this conception of free will arose within a context of *unrequited blame*. That is, when the slaves protested the nobles' treatment of them, and the nobles did not modify their behavior by feeling guilty, this dynamic produced feelings of *ressentiment*, which then led to the belief in libertarian free will as a means of expressing it. Of course, it is entirely consistent with this that libertarian free will is *not* a condition of blame or responsibility, because it does not follow from this that *ressentiment* is essential to blame or responsibility. The kind of blame, free will, and responsibility Nietzsche wants to "overcome" certainly are motivated by *ressentiment*, but this does not mean that *all* forms of blame, free will, and responsibility are.[74]

To make this stronger claim, Nietzsche would not only have to think that responsibility must, in the end, be analyzed in terms of the metaphysical interpretation; it must also be the case that he is categorically against expressions of guilt, remorse, and contrition. He must be against these even when they are expressed in healthy ways as a

[74] See Chapter 4, §4.2. There I argue that the ancient Greeks had an idea of guilt and responsibility that was compatibilist and free of *ressentiment*.

means to improving oneself or restoring one's relations with others. I do not believe that Nietzsche held this stronger view.[75] The *only* form of guilt that he clearly and emphatically rejects is the perverse and "moralized" form of guilt that emerges within Christianity (see GM II: 20-23), which is both an enduring state of the person and one that moreover can never be expiated.

So, if this interpretation is correct, there is no incompatibility between GM II's "long history" of responsibility and the many contexts where Nietzsche denies our being responsible on the basis that we lack libertarian free will. In those contexts he is either giving a diagnosis of *ressentiment* blame, or he was at that time beholden to the metaphysical interpretation of responsibility. In GM II, to the contrary, he "reverses" this interpretation by analyzing responsibility in terms of the *practice* of holding oneself and others responsible, as constituted by the "reactive affects," namely guilt. And I have suggested he did that not with the express aim of vindicating blame and moral responsibility, but because he wanted to *naturalize* our idea of what it means to be "responsible." Finally, as far as the sovereign individual is concerned, if he represents an *ideal* of responsibility, then that ideal would presuppose the more general account of responsible agency offered here. For there is nothing aside from bad conscience in GM II's "long history" that plausibly motivates the distinction between reliability and responsibility.

[75] Nietzsche does want to "take the concept of guilt ... out of the world" (TI VI: 7), but by this he means he wants to restore "innocence" *to the world* by ridding it of Christianity's notion of guilt, which is both a perversion of subjective guilt and requires belief in a "moral world order."

References

Works by Nietzsche

A: *The Antichrist*. 1888. In W. Kaufmann, trans. and ed., *The Portable Nietzsche*. New York: Viking Press, 1954

BGE: *Beyond Good and Evil*. 1886. In W. Kaufmann, trans. and ed., *The Basic Writings of Nietzsche*. New York: Modern Library Edition, 2000.

D: *Daybreak*. 1881. R.J. Hollingdale, trans., Maudemarie Clark and Brian Leiter, ed. New York: Cambridge University Press, 1997.

HA: *Human, All Too Human*. 1878. R.J. Hollingdale, trans. Cambridge: Cambridge University Press, 1996.

GM: *On the Genealogy of Morality*. 1887. Clark, Maudemarie, and Swenson, Alan J, t rans. Indianapolis: Hackett Publishing Company, 1998.

TI: *Twilight of the Idols*. 1888. In W. Kaufmann, trans. and ed., *The Portable Nietzsche*. New York: Viking Press, 1954

Other Works

Acampora, Christa. 2006. "On Sovereignty and Overhumanity: Why it Matters How We Read Nietzsche's Genealogy II, 2*," International Studies in Philosophy* 36: 127-45.

Anderson, R. Lanier. 2013. *Nietzsche on Autonomy*. In Gemes and Richardson, ed., *The Oxford Handbook on Nietzsche*. Oxford: Oxford University Press.

Clark, Maudemarie. 2015a. "Nietzsche's Contribution to Ethics." In *Nietzsche on Ethics and Politics*, 62-74. Oxford: Oxford University Press.

___. 2015b. "Nietzsche on Free Will, Causality, and Responsibility." In *Nietzsche on Ethics and Politics*, 75-96. Oxford: Oxford University Press.

Dannenberg, Jorah. 2017. "Promising by Right." *Philosopher's Imprint* 17(22): 1-18.

Gemes, Ken. 2006. "Nietzsche on Free Will, Autonomy, and the Sovereign Individual." *Proceedings of the Aristotelian Society, Supplementary Volumes,* 80, 321-38.

Hieronymi, Pamela. Forthcoming. "Freedom, Resentment, and the Metaphysics of Morals." Available at: https://ucla.app.box.com/v/HieronymionStrawson

Hatab, Lawrence J. 2009. "Breaking the Contract Theory: the Individual and the Law in Nietzsche's Genealogy." In Siemens, H.W., and Roodt, V., eds., *Nietzsche, Power, and Politics: Rethinking Nietzsche's Legacy for Political Thought*. Berlin: W. de Gruyter, pp. 169-88.

Janaway, Christopher. 2007. *Beyond Selflessness*. Oxford: Oxford University Press.

Leiter, Brian. 2011. "Who is the 'Sovereign Individual'? Nietzsche on Freedom." In Simon May ed., *"Nietzsche's On the Genealogy of Morality": A Critical Guide*.

Cambridge: Cambridge University Press.
__. 2007. “Nietzsche’s Theory of the Will.” *Philosopher’s Imprint* 7 (7): 1-15.

McKenna, M. and Pereboom, D. 2016. *Free Will: A Contemporary Introduction*. New York: Routledge.

Migotti, Mark. 2013. “A Promise Made is a Debt Unpaid: Nietzsche on the Morality of Commitment and the Commitments of Morality.” In Gemes and Richardson, ed., *The Oxford Handbook on Nietzsche*. Oxford: Oxford University Press.

Poellner, Peter. 2009. “Nietzschean Freedom.” In Gemes and May, ed., *Nietzsche on Freedom and Autonomy*. Oxford: Oxford University Press.

Reginster, Bernard. 2017. “What is the Structure of Genealogy of Morality II?” *Inquiry* 61 (1), 1-20.
__. 2011. “The Genealogy of Guilt.” In May, Simon ed., *Nietzsche’s “On the Genealogy of Morality”: A Critical Guide*. Cambridge: Cambridge University Press, 56-77.

Rutherford, Donald. 2011. "Freedom as a Philosophical Ideal: Nietzsche and His Antecedents." *Inquiry* 54(5): 512-40.

Russell, Paul. 2011. “Moral Sense and the Foundations of Responsibility.” In R. Kane ed., *The Oxford Handbook of Free Will: Second Edition*. Oxford: Oxford University Press, 199-220.
__. 1995. *Freedom and Moral Sentiment*. New York: Oxford University Press

Snelson, Avery. 2019. “Nietzsche on the Origin of Conscience and Obligation.” *Journal of Nietzsche Studies* 50(2): 310-331.

Strawson, P.F. 1962. “Freedom and Resentment.” In G. Watson ed., *Free Will: Second Edition*. Oxford: Oxford University Press, pp. 72-93.

Todd, Patrick. 2016. “Strawson, Moral Responsibility, and the Order of Explanation: An Intervention.” *Ethics* 127: 208-240.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

Watson, Gary. 1996. “Two Faces of Responsibility.” *Philosophical Topics* 24(2): 227-48.
__. 1987. “Responsibility and the Limits of Evil: Variations on a Strawsonian Theme.” In Watson ed., *Agency and Answerability*. New York: Oxford University Press.

Zamosc, Gabriel. 2011. “The Relation Between Sovereignty and Guilt in Nietzsche’s *Genealogy*.” *European Journal of Philosophy* 20: 107-142.

Part II

Conscience, Guilt, and Trust

Chapter 3

Nietzsche on the Origin of Conscience and Obligation

Abstract: The second essay of Nietzsche's *Genealogy of Morality* offers a naturalistic account of the development of conscience, understood to be a faculty uniquely responsive to remembering and honoring obligations. This paper attempts to solve an interpretive puzzle that is invited by the second essay's explanation of the emergence of non-moral obligation. Ostensibly, Nietzsche argues that the conscience and our concept of obligation originated within contractual ("creditor-debtor") relations, when creditors punished delinquent debtors (GM II: 5). However, this interpretation, which I call the *contractualist reading*, is incoherent and subject to an insoluble bootstrapping problem. I argue instead that Nietzsche provides two accounts of non-moral obligation in the second essay, and that the conscience originated in the morality of custom to track rule-prohibitions ("I will nots" [GM II: 3]). Appealing to primatologist Frans de Waal's work, I argue that Nietzsche conceives of these "I will nots" as *prescriptive rules*, and hence as involuntary and reciprocal obligations that, unlike contractual debts, do not require the making of promises.

Keywords: conscience; obligation; contractualist reading; rules; debts

I. Introduction

The second essay of the *Genealogy of Morality* is an ambitious text that is structured around explaining the origins of bad conscience and guilt, at which point the conscience becomes a *moral* faculty, as we typically think of it today. However, prior to the emergence of bad conscience, the essay provides a history of human punishment, socialization, and profiles the emergence of non-moral conscience. Like the essay as a whole, Nietzsche's developmental account of the conscience is fragmentary and immensely complex; it takes on numerous forms and goes through various stages of development throughout the essay, in response to different environmental pressures. As Nietzsche says, the conscience "has behind it a long history and metamorphosis" (GM II: 3).

My focus here will be on the most incipient, non-moral form of conscience Nietzsche discusses, which predates the ability to feel guilt and bad conscience.

Specifically, my aim is to reconstruct the conditions in which this earliest form of conscience would have originated, and to describe in some detail the nature of the obligations it developed to track. In the broadest sense, the conscience is understood by Nietzsche to be a "consciousness of" one's obligations, or a kind of “memory” (GM II: 1, 3, 5), which emerges within contexts of pre-existing normative practices. A background of norms and corresponding expectations is presumed by Nietzsche because this "memory" develops as a result of punishment. Punishment, on his view, is both a mechanism for *enforcing* rules or norms, and also a "mnemo-technique" (GM II: 3), a procedure for *creating a memory* of those rules or norms. In short, the conscience initially develops so that we can remain vigilant of the expectations of others, so that we can avoid punishment. But under what conditions were we punished? That is, under what conditions did the conscience and our concept of obligation originate?

Ostensibly, Nietzsche thinks the conscience originated within the context of contractual relationships, or what he calls “creditor-debtor” (GM II: 5) relationships, to ensure that we keep our promises to one another. I call this the *contractualist reading*. On this interpretation, the conscience originated to track obligations qua debts, which we acquired as a result of making promises of repayment. Though I do not deny that this interpretation has a clear basis in Nietzsche's text, I argue in Section 2 that it is incoherent and subject to an insoluble bootstrapping problem. It requires that the debtor be able to communicate promises as a condition of *forming* contractual relationships, as argued by Bernard Reginster (2011, 2017), but contractual relationships are also necessary to explain the *ability* to communicate promises, since doing so presumes the existence of conscience. Accordingly, in the remainder of the paper I

develop an alternative interpretation to the contractualist reading, which I call the *rules reading*.

I develop this alternative through a close analysis of the third aphorism, where Nietzsche appeals to the conscience to explain how humans became reliable, or "regular" and "predictable" in their behavior, which he claims is a "presupposition" of the ability to make promises (GM II: 1, 2). We became reliable on his account by learning to conform our behavior to rules, “I will nots” (GM II: 3). He moreover takes these "I will nots" to be more basic than the idea of debt, describing them as "primitive requirements of social co-existence" (GM II: 3). In Section 3, I explain these remarks by showing that Nietzsche conceives of “I will nots” as a species of what primatologist Frans de Waal has called *prescriptive rules*, which are rules imposed by agents on other agents within dominance hierarchies. On the rules reading, Nietzsche conceives of the conscience, at this early stage in our moral development, as conferring only an ability to remain conscious of social expectations generally, and to conform our behavior to such expectations.

Finally, in Section 4, I argue that Nietzsche conceives of these "I will nots" as *involuntary* or *reciprocal* obligations. Though they are "connected," as he says, to a promise one has made to follow the rules, in order to "live within the advantages of society" (GM II: 3), this promise is only *implicit*. Consequently, these "I will nots" are not acquired as a result of making promises, and therefore do not presume the existence of conscience. Thus, I aim to show that Nietzsche offers a genealogy of *two* forms of non-moral obligation in the second essay—involuntary rules and contractual debts—and since he takes the former to predate the latter, he is able to avoid the bootstrapping problem that plagues the contractualist reading.

II. The Contractualist Reading

According to the contractualist reading, the conscience was “bred” in humans unwittingly in response to the need to be able to keep promises, subsequent to forming contractual or "creditor-debtor" relationships (GM II: 5). As Bernard Reginster has noted, “contractual relationships are established by *promising*, and so they involve the whole apparatus designed to make promising possible, particularly the recourse to the infliction of pain” (2017: 4; 2011: 59). Here Reginster is using “promising” in two different senses. In the first instance he has in mind the act of *communicating* a promise, and in the second the *practice* of promising, specifically as this relates to the ability to *keep* or sustain the motivation to fulfill one's promise. This latter ability, on his account, is what the development of conscience enables—that is why it was "bred" in human beings (GM II: 1). As Reginster elaborates, "The purpose of the necessary breeding is thus to ensure not just the memory that a promise was made, but also the persistence of the motivation to keep it” (2017: 4; 2011: 58, 75).

On this view, contractual relationships are instrumental to the development of conscience, because they are created when one person, the debtor, makes a promise to repay another, the creditor, for some good or service. Subsequent to making this promise, the creditor then held the debtor to that expectation of repayment, but since the debtor lacked a conscience, and therefore a “memory” of his obligation, he failed to repay and was punished by the creditor as an alternative means of compensation. This punishment, Nietzsche insists, was not intended to reform the debtor, teach him a lesson, or elicit feelings of guilt (GM II: 14). It was administered simply because it was pleasurable, because it gratified the creditor’s “instinct for cruelty” (GM II: 5), and therefore provided him with an alternative means of repayment or satisfaction.

However, because punishment was also a "mnemo-technique" (GM II: 3), a procedure for creating memory, the creditor's punishment had the unintended effect of producing in the debtor a conscience, a faculty which allowed him to both "remember" his debt and sustain the motivation to keep his promise.[76]

Two considerations support the idea that Nietzsche endorses the contractualist reading. For one thing, he takes the last and "highest form" of conscience (GM II: 3), as instantiated in the "sovereign individual," to underwrite what he calls "permitted promising" (GM II: 2). This is plausibly interpreted as a metaphor for *trust*, and so the conscience is described by Nietzsche as emerging along this developmental path that ultimately enables and justifies trust of others, which of course presumes that we first make promises. Secondly, and more importantly, Nietzsche seems to endorse the contractualist reading explicitly in the following quote.

> Calling to mind these contract relationships admittedly awakens various kinds of suspicion and resistance toward the earlier humanity that created or permitted them ... Precisely here there are *promises* made; precisely here it is a matter of *making* a memory for the one who promises ... In order to instill trust in his promise of repayment, to provide a guarantee for the seriousness and the sacredness of his promise, to impress repayment on his conscience as a duty, as an obligation, the debtor—by virtue of a contract—pledges to the creditor in the case of non-payment something else that he "possesses," over which he still has power, for example his body or his wife or his freedom or even his life ... Above all, however, the creditor could subject the body of the debtor to all manner of ignominy and torture .... (GM II: 5)

Nietzsche's argument here seems to be that the act of *making* promises, because doing so created a dynamic in which creditors punished debtors and produced the conscience,

---

[76] In his more recent article, Reginster attributes the development of conscience to the morality of custom. However, he also says it emerged along with "promise keeping," and so it would seem that contractual relationships are still operating in the background (2017: 4). Similarly, Aaron Ridley (2009: 182) holds that humans became reliable by making promises. As will become clearer in Section 3, I think Reginster and Ridley here make the mistake of conceiving of reliability of behavior as a condition of *keeping* a promise, when it is also and more basically a condition of being able to *make* one.

made it possible to *keep* promises, just as Reginster's remarks above indicate. However, a consequence of this argument is that debtors first had to make promises of repayment they did not fulfill, prior to their being punished, as a condition of *forming* contractual relationships.

For this reason, as I will now show, Nietzsche's explanation of origin of conscience and obligation is incoherent and self-undermining on Reginster's interpretation, because he conceives of the ability to communicate promises as a *condition* of forming contractual relationships, and contractual relationships are moreover necessary to explain the *ability* to communicate promises. First, note that promising is not simply the act of communicating an intention. As Reginster says, "to make a promise is to commit to doing something at some appointed future time even if doing so has by then become contrary to my 'private desires and advantages' (D 9)" (2017: 4). Promising involves expressing a commitment that is in a key sense *binding*, and that is what the debtor does in the above context. By making a promise, repayment is conveyed by the debtor and understood by both him and the creditor as something *non-optional*, as a kind of *requirement*. This is because promising, unlike merely communicating an intention, involves communicating an *obligation*. Indeed, to make a promise just is to "communicate an intention to undertake an obligation."[77] Thus, to communicate a promise, the debtor must *already* have a concept of obligation.

I will have more to say about obligation in the next section, but here it bears emphasizing that we need not build too much into the idea. In particular, it need not be the case that the debtor understands himself to have a *moral* obligation to repay his creditor, or that he would feel guilty if he failed to do so. It need only be the case that the

[77] See Owens (2006: 54); Raz (1977: 218); Watson (2009: 156).

debtor recognizes that he is in a minimal sense bound to repay him: that he "should" or "ought" to do so, as Reginster says above, quoting Nietzsche, regardless of his "private desires and advantages" (D 9). Obligations have this character because they are *social requirements*, actions which I am bound ("should" or "ought") to perform by others regardless of my personal desires.[78] To be slightly more technical, the mental state of obligation, its unique sense of "ought" or "should," is a consideration given deliberative priority in order to secure reliability of behavior,[79] which, when acted upon, has the social function of satisfying the expectations of those who have power or authority. Communicating an obligation, then, requires the ability to discriminate between those courses of action that are purely personal—"hypothetical imperatives," as Kant would characterize them—and those that are non-hypothetical.[80] That is, promising requires that I be capable of recognizing and conveying the course of action as making a claim on me even if I am not inclined to do it, as having a special kind of priority given its social significance, as something that is recognized by me as required or "obligatory" in the sense that its performance is not solely up to me.

Above, the debtor's promise has this non-hypothetical character. This promise is made in a social dynamic in which the creditor will hold him to an expectation of repayment *regardless* of whether he wants to do so at the appointed time, and it is further assumed by Nietzsche that the creditor has the *authority* to hold the debtor accountable, should he fail to do so. As Nietzsche claims, this dynamic establishes that the creditor has a "directive and right to cruelty" (GM II: 5). Consequently, the idea that

---

[78] For more on obligation as a social requirement, see Feinberg (1970: 244); Foot (1972: 308); and Strawson (1961: 5))

[79] This is Bernard Williams' definition of obligation. See Williams (1985: 185).

[80] Hypothetical imperatives recommend the means necessary "to achieving something else that one wants" (Kant 2012: 28), or what one "ought" to do to be effective in satisfying one's desires.

the debtor could make a promise in this context, without having a familiarity with obligations as social requirements, is simply *incoherent*.

Not only this, the debtor also must have a conscience to *make* promises. If the debtor had no conscience, a condition of obligation—specifically, the ability to be *aware* of or have a "consciousness of" obligation—has not been realized. In such a world, the debtor would lack a "memory," not just of *his* obligation or debt, but of the very *idea* of obligation. The contractualist reading, because it maintains that the conscience and our concept of obligation originated as *consequences* of promising, gives rise to the following causal dilemma. It requires the truth of both of the following propositions, but these cannot be held consistently:

> (P1) contractual relationships are formed by communicating promises ("precisely here there are *promises* made" [GM II: 5]), and
>
> (P2) the conscience and our idea of obligation originated as debt within contractual relationships.

If the idea of obligation originates as that of debt, the debtor must be capable of entering into contractual relationships, according to P2. But in order to enter into contractual relationships, the debtor must have some idea of obligation to communicate promises, according to P1. And since contractual relationships are formed by communicating intentions to undertake obligations of repayment (P1), and because the debtor has no concept of obligation prior to entering contractual relationships (P2), this means it would have been *impossible* to form contractual relationships. Consequently, Nietzsche's account of the origin of conscience and obligation is subject to an insoluble bootstrapping problem on the contractualist reading.

I do think Nietzsche has the resources to avoid this problem, though doing so will require rejecting either P1 or P2, and thus rejecting the contractualist reading. What Nietzsche describes in the fifth aphorism is actually a fairly complex transaction, one that is created by the explicit exchange of voluntary obligations and presumes, among other things, a cultural background in which laws, property, and money already exist. Though few scholars have taken note of it, he even assumes the debtor already has a conscience.[81] For these reasons, we might think that what he describes in this passage is a later development of a more rudimentary practice of "proto-promising," one that does not presume the debtor has a concept of obligation or a conscience. I believe this to be the case, which will require that we reject P2 and affirm P1.

Consider, first, why Nietzsche ought to affirm P1. For one thing, he simply takes it for granted that promising is a condition of forming contractual relationships ("precisely here there are promises made"). Secondly, Nietzsche believes contractual relationships involve the kind of transfer of power or rights often thought to coincide with making promises. Specifically, they involve a transfer of property (collateral) and the "directive and right to cruelty" (GM II: 5). This is why he claims contractual relationships coincide with the appearance of "legal subjects" (GM II: 4) and "the most rudimentary form of personal legal rights" (GM II: 8). Despite their "rudimentary" nature, these are complex ideas, the likes of which cannot be found in the animal

[81] Note, the conscience is invoked to explain how the debtor "impress[es] repayment on his conscience as a duty, as an obligation" (GM II: 5). Simon May is the only scholar I am aware of who is brought attention to this peculiar feature of the passage. On his reading, Nietzsche simply takes it for granted that the debtor has a "strong" and "ethically charged notion of personal accountability" (May 1999: 56). I have reservations about May's account, since it would seem to assume that at this stage in Nietzsche' story the debtor is already capable of feeling guilt, the "feeling of personal obligation" (GM II: 8), which creditor-debtor relationships are supposed to explain. That said, I agree with May that the debtor must possess a minimal sense of obligation, as articulated above. I do not take it to be particularly "strong" or "ethically charged," though, for reasons that will become apparent in Section 3.

kingdom, ideas which only serve to reinforce Nietzsche's claim that creditor-debtor relationships represent "man's preeminence with respect to other creatures" (GM II: 8).

Finally, I will argue in Section 4 that we ought to interpret Nietzsche as rejecting P2 because "proto-promising" just is the practice of reciprocity, and Nietzsche's account of reciprocal obligation is offered in the third aphorism, in connection to the "I will nots." Consequently, I will argue that Nietzsche rejects P2 because he recognizes the existence of non-moral obligations that are more basic than, and indeed preliminary to, voluntary debts. I will now turn to explaining the origin and nature of these "I will nots," with an eye toward showing how they constitute a more primitive form of obligation.

III. The Rules Reading

Early in the second essay, Nietzsche observes that being reliable, or "regular" and "predictable" (GM II: 1, 2) in one's behavior, is a presupposition of promising, of being able to "vouch for [oneself] *as future*" (GM II: 1). This is because, as we have seen, a person who promises commits herself to some future action regardless of her "private desires and advantages" (D 9), which requires an ability to discriminate the obligatory from the non-obligatory, an "ought" which is merely prudential or hypothetical from an "ought" that is "non-hypothetical," as I said above. The idea underlying these remarks is that agents who are "regular" and "predictable" in their behavior are so because they do what is *expected* of them, because they conform their behavior to compulsory norms or rules, rather than act on their strongest desire or whim of the moment. Nietzsche's account of the conscience in the third aphorism is offered to explain how humans became reliable in these ways. There he conceives of the conscience as a kind of *social memory* that made it possible to follow rules and live with others.

> "How does one make a memory for the human animal? How does one impress something on this partly dull, partly scattered momentary understanding, this forgetfulness in the flesh, so that it remains present?" … As one can imagine, the answers and means used to solve this age-old problem were not exactly delicate; there is perhaps nothing more terrible and more uncanny in all of man's prehistory than his *mnemo-technique*. "One burns something in so that it remains in one's memory: only what does not cease *to give pain* remains in one's memory"—that is a first principle from the most ancient (unfortunately also longest) psychology on earth … The worse humanity was "at memory" the more terrible is the appearance of its practices; the harshness of penal codes in particular provides a measuring stick for the amount of effort it took to achieve victory over forgetfulness and to keep a few primitive requirements of social co-existence *present* for these slaves of momentary affect and desire … With the help of such images and processes one finally retains in memory five, six, "I will nots" in connection with which one has given one's *promise* to live within the advantages of society. (GM II: 3)

I will address Nietzsche's conclusion that the "I will nots" are "connected" with a "promise to live within the advantages of society" in the next section.[82] Here I will be focusing on the formation and character of this "memory." My aim is to show that the "I will nots" are basic to cooperative sociality, and as such explain the origin of obligation.

As John Richardson has remarked, "this memory is simply the ability to 'remember' social rules or practices, to be bound by them" (2004: 89). The conscience here consists simply in being able to remember a handful of customary prohibitions, rules like "I will not steal," "I will not kill," and "I will not lie," which Nietzsche considers to be necessary for the maintenance of social life. In fact, Nietzsche claims this memory-making technique belongs to the "longest" and "most ancient psychology on earth." He attributes this "enormous work" to the "morality of custom," claiming it was the "true work of man on himself for the longest part of the duration of the human race, his entire

[82] I assume this is why Reginster and Ridley take reliability to be a *consequence* of promising, rather than a condition of making one. The burden is on me, then, to show that promising is only implicit in this context.

prehistoric work" (GM II: 2).[83] These remarks suggest that the above passage is concerned with some amorphous stage in our evolutionary past, perhaps as far back as the appearance of *Homo*, what Nietzsche above calls the "human animal."[84] Speculation aside, that Nietzsche intends for this form of conscience to extend very far back in our evolutionary past cannot be questioned: the capacity that it is invoked to explain—the ability to follow rules and be reliable in one's behavior—is not distinctly human.[85]

As Frans de Waal remarks, "All animals conform to social rules. That is, their conduct toward conspecifics is to some degree predictable" (1991: 337). Social rules, so understood, are regularities that circumscribe behavior, because they are imposed on agents by other agents. De Wall refers to such rules as *prescriptive rules* (1996: 96). As we can see, this is the kind of rule that Nietzsche describes above. De Waal moreover believes that prescriptive rules generate obligations, or possess an "ought quality," because they are learned behavioral patterns "actively upheld through reward and

---

[83] Nietzsche does not tell us the origin of the morality of custom, only that it is "prehistoric" (GM II: 2) and goes back "many millennia" (D 14). I agree with Iain Morrisson, (2018: 986, fn.7) that prehistoric humans were social and their interactions with one another regulated by customs. To the contrary, a number of scholars take Nietzsche to be committed to a pre-social state of nature, prior to the formation of states (GM II: 17). (See Reginster [2011: 62]; Ridley [1998: 18-19]; Risse [2001: 57]). On Nietzsche's view, the earliest tribes were "organized according to blood-relationships" (GM II: 20), and he believes customs originated as utility rules, "the experiences of men of earlier times as to what they supposed useful and harmful" (D 19). These are what he above refers to as "primitive requirements of social co-existence" (GM II: 3). These eventually acquire the status of a custom—*Sitte*—a norm that has been passed down through the generations and become a "*traditional* way of behaving" (D 9). Accordingly, I will rely on a broad understanding of the morality of custom throughout, taking it to be a primitive form of human social organization that enforced rules and secured reliable interactions among group members. It is in this capacity that Nietzsche appeals to the morality of custom in the second and third aphorisms.

[84] Similarly, Brian Leiter claims that the third aphorism describes "a phenomenon of pre-history: we are discussing what the animal man had to be like before regular civilized intercourse with his fellows ('the advantages of society') would even be possible" (2002: 229).

[85] Nietzsche was certainly aware of this, because he attributes this and more complex capacities to animals in *Daybreak* 26. There he claims that animals are capable of "objective awareness," of treating themselves as the object of another's gaze, and thus possess the kind of social awareness involved with having a conscience. Like Darwin (1879/2004: 119-51), Nietzsche believed that human conscience was constructed from various capacities already present in non-human animals.

punishment." In other words, obligation emerges with the thought, or it would be more accurate to say, the *sense* that: "I better do X or suffer negative consequences." This is an inference that the formation of conscience, on Nietzsche's view, allows us to remain *aware* of. Finally, according to de Waal, prescriptive rules are made possible by hierarchical relationships in which A (typically a dominant) holds another B (typically a subordinate) to an expectation of conformity, which B (and others) learn by being punished (1991: 340; 1996: 92).[86] (We will see evidence in a moment to suggest that Nietzsche also thinks the conscience developed within dominance hierarchies.) De Waal concludes, "a prescriptive rule is born when members of the group learn to recognize the contingencies between their behavior and that of [others] and act so as to minimize negative consequences."[87]

Ok, but isn't the force of such imperatives still merely *hypothetical*? After all, according to de Waal, obligations are generated when agents with power or authority routinely hold others to expectations of conformity, and so without the threat of punishment, we would not abide such rules. Kant would have characterized these as imperatives of "prudence," because conformity to them is still just predicated on avoiding punishment, and therefore merely a "means to another purpose" (Kant 1785/2012: 29), namely, avoiding suffering for the sake of the individual's long-term welfare. However, Nietzsche evidently meant *not* to say this. As he claims in *Daybreak*, conformity to customs requires obeying the rule "*not* on account of the useful consequences it may have for the individual, but so that the hegemony of custom,

[86] See also Clutton-Brock and G.A. Parker (1995: 211).

[87] It is important to note, at least in primates, conformity to a prescriptive rule is not just an instinctual response. De Waal believes they have an *awareness* of prescriptive rules, which is conspicuous both in their efforts to avoid detection, when violating norms, and when "tattling" on one another. For some rather comical anecdotes, see de Waal (1991: 338-9). See also Boehm (2012: 106).

tradition, shall be made evident *in despite* of the private desires and advantages of the individual" (D 9, emphasis mine).

To make sense of Nietzsche's claim, I think we must consider another aspect of obligation. That is, in order to understand what is peculiar about the *motivation* to conform our behavior to obligations, on his view, we must also consider why doing so is of vital importance. As H.L.A. Hart observes, "Rules are conceived and spoken of as imposing obligations … when the general demand for conformity is insistent and the social pressure brought to bear upon those who deviate or threaten to deviate is great" (1961: 86). Above, Nietzsche believes the "I will nots" have this same character. They are understood to be "primitive requirements social co-existence," meaning that conformity to them is a prerequisite of community membership—they must be followed to live amongst others *at all*. Conformity with such rules is thus *obligatory* or *non-optional*; indeed, a condition of cooperative sociality itself. And so while it may be the case that such rules would not be obeyed if we weren't threatened with punishment, it's *also* the case that they must *always* be obeyed, because it is always being expected of us that we will obey them. In other words, obligation not only involves a way of relating to an action and being motivated to act; obligation is also an omnipresent *social fact*, because certain actions must be performed and others avoided, simply in order to live with others at all.

The conscience is invoked by Nietzsche in the third aphorism to explain how we became "regular" and "predictable" in our behavior, and so to explain how this strictly third-personal aspect of obligation became internalized and the human animal self-regulating, by coming to see these social requirements in a particular way. As Reginster correctly observes (2017: 4), this is essentially a matter of being able to "overcome" or

"disregard" one's "private desires and advantages" (D 9), but above Nietzsche suggests we acquired this ability merely in virtue of the fact that rules must be followed to live with others, not due to the need to be able to keep promises. An "I will not" is understood by Nietzsche to be a *rule,* an action which I am bound ("should" or "ought") to perform by others regardless of my personal desires, the function of which is to secure reliable interactions with others as a necessary condition of cooperative sociality. So, obligation consists, in the first instance, in the *social fact* of obligation, in the fact that others with power or authority hold us to expectations, constituting rules. However, obligation also consists, more interestingly, in a distinct *mental state*, an awareness of a course of action that "ought" to be done regardless of the agent's personal desires.

Bernard Williams provides a helpful characterization of this "ought." He defines an obligation as "a consideration given deliberative priority in order to secure reliability" (1985: 185). On his account, just as on Nietzsche's, the function of obligation is "to secure reliability, a state of affairs in which others can reasonably expect me to behave in some ways and not in others" (1985: 187). Acting on an obligation, then, is not just an instinctual or automatic response; it requires the ability to reflect, to stand back, and to assess an action's social consequences. It requires, as Nietzsche says above, an ability to no longer be a "slave to momentary affect and desire" (GM II: 3). He even thinks this is the origin of reason itself: "With the help of this kind of memory one finally came 'to reason'!—Ah, reason, seriousness, mastery over the affects, this entire gloomy matter called reflection" (GM II: 3).[88] By creating this possibility of deliberative conflict in the

[88] Paul Katsafanas (2016: 89-90) observes that scientists and philosophers in the 19th Century typically distinguished instinctual actions or responses from learned behaviors, taking the former to be unreflective and the latter to require awareness of an end or goal (e.g., avoiding punishment). It is plausible to think Nietzsche had the same distinction in mind in the third aphorism.

human animal, the development of conscience transformed the merely third-personal *social fact* of obligation into a way of *relating* to them. In other words, the conscience gave rise to this peculiar notion of "ought."

We are now in a position to state why the imperative to follow rules or customs is non-hypothetical. The conscience originates as an "inner voice" reminding us to obey these rules, but this voice is of an "ought" that is *unconditional*, because it takes the form of a *command*. As Nietzsche explains in *Beyond Good and Evil:*

> Inasmuch as at all times, as long as there have been human beings, there have also been herds of men (clans, communities, tribes, peoples, states, churches) and always a great many people who obeyed, compared with the small number of those commanding—considering, then, that nothing has been exercised and cultivated better and longer among men so far than obedience—it may be fairly assumed that the need for it is now innate in the average man, as a kind of *formal conscience* that commands: "thou shalt unconditionally do something, unconditionally not do something else," in short, "thou shalt." (BGE 199)

In this quote, Nietzsche offers a description of the inner voice of conscience, a description of its *form*. He tells us that the conscience originated to secure obedience to commands within social dominance hierarchies, which we have lived in since the inception of our species, and so the form this voice takes is that of a *command*: an "unconditional thou shalt."

We can be certain that, by "unconditional," Nietzsche does *not* mean the same thing as what Kant meant. According to Kant, the categorical imperative is "unconditional" in the sense that compliance with it requires "the dissociation from all interest in willing from [from a motive of] duty" (Kant 1785/2012: 44). For Nietzsche, this is a fiction—there is no "pure" moral motive. But he also recognizes, at least since *Daybreak*, that acting on an obligation is different than acting on the basis of hypothetical imperatives. There he appealed to the morality of custom to offer a

naturalistic explanation of the categorical force of moral norms, acknowledging that this cannot consist simply in considerations surrounding what is prudent or "useful" (D 9).[89] Hypothetical imperatives are generated *conditionally* on the basis of the agent's other desires or ends, and so if the agent's desires change or she gives up her end, the "ought" also disappears. As we saw in the case of the debtor's promise, obligations are not dependent on our desires in this way. So, what is essential to the idea of obligation and in need of a naturalistic explanation is this idea of an "ought" which is distinct from the merely prudential "ought," one that is in some sense not dependent on, nor entirely divorced from, the agent's other desires.

It is this gap which the conscience as a social memory is offered to fill-in. The conscience is the inner voice of this non-hypothetical ought. It is non-hypothetical because, as Phillipa Foot observes, "Lacking a connection to the agent's desires or interests, 'should" in this case does not stand 'unsupported and in need of support'; it requires only the backing of the rule" (1972: 309). This does not imply that the imperative can be satisfied in the absence of the agent's other motives (e.g, fear and the drive for self-preservation); it implies only that it does not depend on these for its *existence*. Also, as Nietzsche tells us above, the motive to obey rules requires regarding them not as a means to something else one wants, but as a *command*. A custom is a "higher authority which one obeys, not because it commands what is *useful* to us, but because it *commands*" (D 9). Importantly, this obedience need not be "pure." Commands may be obeyed out of fear, reverence, awe, or a mixture of these and other motives. What it means to obey a command "unconditionally" is simply that one complies with it in recognition of it *as such*, that is, by regarding it as something that

[89] See Clark and Leiter (1997: xxx).

"ought" to be done even when doing so is contrary to one's "private desires and advantages" (D 9). By doing so, one gives this "ought" deliberative priority, and she therefore recognizes it as having different status and significance than the "ought" of hypothetical imperatives.

Finally, I think Nietzsche offers a compelling explanation for how we came to act on the basis of such "oughts." He suggests that our doing so is simply a habit or tendency we have acquired to obey the commands of those who have rank, because we have lived in hierarchical relationships since the inception of our species. He calls this habit or tendency the "herd instinct of obedience" (BGE 199).[90] A command, being a compulsory order from a source of power or authority, when enforced consistently creates a norm or rule, per de Waal's account of prescriptive rules. And so the "herd instinct of obedience" is not merely a tendency to obey those with rank and to regard their commands as "unconditional," but a tendency to conform one's behavior to norms or rules by regarding *them* as "unconditional." We became "regular" and "predictable" in our behavior, then, by developing a conscience and this herd instinct of obedience. If this is right, the distinct "ought" of obligation is the result of this long history of "breeding" within social dominance hierarchies, a habit or tendency we have acquired to cognize some modes of conduct as *social requirements*. The non-hypothetical "ought" of obligation *just is* the voice of the "herd instinct of obedience."

[90] The herd instinct is a disposition to conform one's behavior to that of others (GS 116). The "herd instinct of obedience" is a disposition to give uptake to the commands of those who are perceived to have rank or authority. Stanley Milgram's famous and disturbing "obedience study" illustrates just how ingrained this disposition is in human beings, and how it is distinct from and may conflict with, moral motivation. (Most participants in the study were reluctant to shock the confederate because they felt it was wrong, but were somewhat relieved and more willing to do so when assured by the authority figure that he would take responsibility.) See Milgram, "Behavioral Study of Obedience," *Journal of Abnormal and Social Psychology*, 67: 371-78.

IV. Reciprocity and the Communal Bargain

I will now address Nietzsche's claim that the "I will nots" are connected with a "promise to live within the advantages of society" (GM II: 3). In the third aphorism, Nietzsche describes two practices that are "prehistoric" (GM II: 3) and basic to cooperative sociality—so basic that we see evidence of both in non-human animals—which later came to be *interpreted* in terms of the creditor-debtor schema and the idea of equivalence. These practices are punishment and reciprocity. Unfortunately, space precludes me from getting into the details of Nietzsche's argument here; however, it is sufficient for my purposes to show that "I will nots" and debts are not equivalent on his account, and that the former are involuntary obligations that one acquires without making promises.

Nietzsche describes two kinds of creditor-debtor relationships in the second essay, dyadic contractual agreements between individuals (GM II: 5), and the communal relationship between the individual qua debtor and society qua creditor (GM II: 9). On my reading, Nietzsche takes the communal dynamic described in the third aphorism to be the more fundamental. We saw evidence of this in both the third aphorism, where he claims the "I will nots" are "primitive requirements of social co-existence" (GM II: 3), and in BGE 199, where he claims human beings have always lived in social dominance hierarchies.[91] Subsequent to the formation of dyadic contractual agreements, which arose in concert with "the basic forms of purchase, sale, exchange, trade, and commerce" (GM II: 4), this communal dynamic came to be "interpreted" in terms of two concepts

[91] We also see evidence in GM II: 8, which on the face of it seems to present evidence to the contrary. Specifically, Nietzsche remarks that the economic instruments associated with the creditor-debtor schema, along with the "rudimentary" ideas of "exchange, contract, guilt, right, obligation, compensation" were "*transferred* … onto the coarsest and earliest communal complexes …" (GM II: 8). This presumes that humans were living in groups, prior to the advent of creditor-debtor relations, despite Nietzsche's claim that they are "older even than the beginnings of societal associations and organizational forms" (GM II: 8). Space precludes defense here, but I believe he takes these "associations" and "forms" to be *government* or *political* organizations.

introduced by the creditor-debtor schema: the notion of debt and the principle of equivalence, "the idea that every injury has its *equivalent* in something and can really be paid off" (GM II: 4). This is what Nietzsche provides analysis of in the ninth aphorism.

"Interpretation" is a term of art for Nietzsche. Interpretations are explanations that "integrate" a practice "into a system of purposes" (GM II: 12). That is, an interpretation comes into existence by conceptualizing a pre-existing practice in terms of specific aims or goals, and in doing so infuses what is otherwise a mere practice—a series of procedures, performed routinely—with meaning. Let's first consider punishment. Punishment consists of a "relatively *permanent*" element, "the practice, the act, the 'drama,' a certain strict sequence of procedures," and a "*fluid*" element, "the meaning, the purpose, the expectation tied to the execution of such procedures" (GM II: 13). Nietzsche adds that "the procedure itself will be something older, earlier than its use for punishment, that the latter was first placed into, interpreted into the procedure (which had long existed, but was practiced in another sense)" (GM II: 13). The stable or "permanent" element of punishment, I suggest, is simply that it is a response "to an injury suffered, which is vented on the agent of the injury" (GM II: 4). This would seem to be the form of punishment we find in the third aphorism, which is described as little more than outward-directed anger in response to violating an expectation of conformity. As I argued previously, this dynamic has the effect of creating and enforcing prescriptive rules.[92] In the fifth aphorism, on the other hand, punishment has been interpreted,

[92] As noted by Maudemarie Clark (2015: 93), Nietzsche appeals to a similar notion of communal "punishment" in the ninth aphorism, prior to the idea of equivalence. He also acknowledges that "punishment" is not uniquely human: "Seeing-suffer feels good, making-suffer even more so—that is a hard proposition, but a central one, an old powerful human-all-too-human proposition, to which, by the way, even the apes might subscribe: for it is said that in thinking up bizarre cruelties they already abundantly herald and, as it were, 'prelude' man" (GM II: 6).

"integrated into a system of purposes." Specifically, it has become a method for paying off debts, in accordance with the principle of equivalence.

"What is the difference between a mere obligation, a sense that one ought to behave in a certain way, or even that one owes something to someone, and a debt, properly speaking? The answer is simple: money. The difference between a debt and an obligation is that a debt can be precisely quantified" (Graeber 2011: 21). As David Graeber here observes, a debt is an obligation that has a *valuation* attached to it, and this is a point which Nietzsche himself stresses in the fifth aphorism. Creditors administered punishment in a manner that seemed to them "commensurate to the magnitude of the debt," and "exact assessments of value developed from this viewpoint, some going horribly into the smallest details—*legally* established assessments of the individual limbs and areas on the body" (GM II: 5). This idea of equivalence is completely absent from the third aphorism because it describes an *earlier* practice of punishment, one that makes viable the venture of cooperative sociality but is not yet connected to ideas of justice and fairness.[93] So the "I will nots" are not conceived by Nietzsche *as debts*. They represent a more primitive form of obligation—rules.

Now let's consider reciprocity. The fundamental basis of any social venture that aims to secure a common goal, among human or non-human animals, is cooperation. However, to cooperate simply means "to act together," and so cooperation covers a wide swath of collaborative activities, ranging from mutualism, a form of "acting together" in which A and B benefit simultaneously (e.g., the coordinated hunting efforts of pack animals), to contractual relations, which are a uniquely human form of cooperation in which A provides some good or service to B on the condition that B promises to

[93] Equivalence is "the oldest and most naïve canon of moral justice" (GM II: 8).

reciprocate (Rothstein and Pierotti 1988: 198). Reciprocity is a form of cooperation that lies between mutualism and contracts. To reciprocate means to "give and take mutually" or "to make a return for something" (Rothstein and Pierotti 1988: 198). Like the forming of contracts, reciprocity involves a *conditional* exchange of favors, but like mutualism it does not require the making of promises. De Waal, appealing to Robert Trivers (1972) theory of reciprocal altruism, defines reciprocity as an exchange of favors in which:

(1) The initial act, while beneficial to the recipient, is costly to the performer,

(2) there is a time lag between giving and receiving, and

(3) giving is contingent on receiving.

Since giving is contingent upon receiving, reciprocal relations have the same underlying structure as contractual relationships. In fact, de Waal often describes them as quasi-contractual relations governed by implicit promises.[94]

What Nietzsche is describing in the third aphorism is a reciprocal, not a contractual, relationship. Recall, the "I will nots" are "connected" to a promise "one has made … in order to live within the advantages of society." First, why is the promise "connected" to these "advantages?" The answer is that rule-following is a *condition* of receiving its protection. As Nietzsche later observes, the community member "lives protected, shielded, in peace and trust, free from care with regard to certain injuries and hostilities to which the human *outside*, the 'outlaw,' is exposed," and in view of which one has "pledged and obligated oneself to the community" (GM II: 9). So, in other words, if the community member does not follow the rules, the community will withdraw its protection, i.e., he will be liable to punishment. As Maudemarie Clark has noted in

[94] See, e.g., Flack and de Waal (2000: 19-20).

reference to this passage, “If you accept the advantages of community life, you are in effect making a bargain with the community, agreeing to go along with the rules that make community life possible” (1994: 36).

However, and secondly, this communal bargain does not imply that society’s protection was offered on the condition that the individual *made* a promise to follow the rules. Just like other social animals who reap the benefits of cooperative sociality without making promises, primitive humans enjoyed the benefits of communal life simply by being born into a community, and they received its protection so long as they followed the rules—even if they never made a promise to do so. As we saw previously, to receive the group’s protection, all that one must do is abide the “primitive requirements of social co-existence” (GM II: 3) that make communal life possible. The individual need not, in addition to this, make a *promise* to follow them. But why, then, does Nietzsche say the rules are “connected” with a “promise one has made?” He says this because one is *obligated* to follow these rules, and so it is *as if* one has made a promise to follow them. In other words, the “I will nots” are *reciprocal* obligations.

Thirdly, and finally, we can be assured that the promise is *only* implicit in this context, because, as I argued previously, what punishment is making the agent "conscious of" is the fact that she is *subject to rules*, or that certain actions are obligatory. In other words, punishment is acquainting her with the basic and indelible reality of the *social fact* of obligation, regardless of whether she ever consented to follow these rules. For this reason, "I will nots" are *involuntary* obligations. Dyadic contractual relationships, to the contrary, are established by making promises (“Precisely here there are *promises* made” [GM II: 5]), and so are created by *voluntary* obligations. In this situation a promise *must* be made, unlike in the third aphorism, because promising is a

condition of receipt: the creditor extends the initial good or service *only if* the debtor produces an expectation of repayment by communicating it. Consequently, it is the making of a promise that gets this whole transaction off the ground in the first place, and so the practice assumes that the promisor already has a concept of obligation and a conscience—just as Nietzsche acknowledges in the fifth aphorism.

If this is right, the third aphorism describes a practice of "proto-promising" that must predate the origin of creditor-debtor relationships, a practice we see evidence of in non-human animals. The "I will not" is *not* acquired subsequent to making a promise; it is an obligation one incurs merely in virtue of living with others and accepting the "advantages" of communal life. Since the "I will nots" are "requirement[s] of social co-existence" (GM II: 3), it makes sense to say they are "connected" to a promise. However, this means only that promising is conceived by Nietzsche as a structural or formal feature of the communal relationship, not a condition of forming it. That said, this reciprocal relationship is eventually *interpreted* as a creditor-debtor relationship on his analysis, but only subsequent to the development of the ideas of debt and equivalence. Consequently, the "I will not" is a *reciprocal* and *involuntary* obligation that, unlike contractual debt, does not presume the ability to make promises.

V. Conclusion

This investigation began by raising awareness of a causal dilemma generated by a natural interpretation of the second essay's account of the emergence of conscience and obligation. According to this "contractualist reading," as I called it, Nietzsche takes promising to be a condition of forming contractual relationships, and such relationships are moreover necessary to explain the ability to make promises. I have instead tried to show that he is not committed to this interpretation, since the third and fifth aphorisms

articulate two different conceptions of non-moral obligation. The third aphorism is offered to explain how human beings became “regular” and “predictable” in their behavior by becoming aware of and conforming their behavior to rules, understood to be obligations that one acquires *involuntarily* merely in virtue of living a social form of life. Dyadic contractual agreements, on the other hand, are created on the basis of *voluntary* obligations, by the making of promises, presuming that the debtor already has a conscience and a concept of obligation.

Relative to the scope of the second essay as a whole, my aims here have been quite modest, but I hope they have not been insignificant. I have tried to show that Nietzsche provides a plausible and naturalistic account of the origins of obligation by offering a genealogy of conscience, understanding it initially to be a faculty that was “bred” in human beings through punishment. On the rules reading, the conscience emerged as the inner voice of a non-hypothetical “ought,” to track rule-prohibitions during the morality of custom, prior to the advent of creditor-debtor relationships. Also, I hope to have gone some way in making sense of the "long history and metamorphosis" (GM II: 3) the conscience went through prior to its becoming a moral faculty, the "bad conscience,” "the consciousness of guilt" (GM II: 4). If the preceding analysis has been right, this history begins as a history of *involuntary* obligations, which humans acquired merely in virtue of being the social creatures we are.

References

Works by Nietzsche

BGE: *Beyond Good and Evil*. 1886. In W. Kaufmann, trans. and ed., *The Basic Writings of Nietzsche*. New York: Modern Library Edition, 2000.

D: *Daybreak*. 1881. R.J. Hollingdale, trans., Maudemarie Clark and Brian Leiter, ed. New York: Cambridge University Press, 1997.

GM: *On the Genealogy of Morality*. 1887. Clark, Maudemarie, and Swenson, Alan J, trans. Indianapolis: Hackett Publishing Company, 1998.

GS: *The Gay Science*. 1882/1887. W. Kaufman, trans. New York: Vintage Books, 1974.

HA: *Human, All Too Human*. 1878. R.J. Hollingdale, trans. Cambridge: Cambridge University Press, 1996.

Other Works

Boehm, Chistopher. 2012. *Moral Origins*. New York: Basic Books.

Clark, Maudemarie. 2015. "Nietzsche on Free Will, Causality, and Responsibility." In *Nietzsche on Ethics and Politics*, 75-96. Oxford: Oxford University Press.

___. 1994. "Nietzsche's Immoralism and the Concept of Morality." In *Nietzsche on Ethics and Politics*, 23-40. Oxford: Oxford University Press.

Clark, Maudemarie and Leiter, Brian. 1998. "Introduction." In *Daybreak*, R.J. Hollingdale, trans., Maudemarie Clark and Brian Leiter, ed. New York: Cambridge University Press.

Clutton-Brock, T.H., and Parker, G.A. 1995. "Punishment in Animal Societies." *Nature* 373: 209-216.

Darwin, Charles. 1879/2004. *The Descent of Man*. London: Penguin Books.

De Waal, Frans. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge: Harvard University Press.

___. 1991. "The Chimpanzee's Sense of Social Regularity and Its Relation to the Human Sense of Justice." *American Behavioral Scientist*, *34* (3), 335–349.

Graeber, David. 2011. *Debt: The First 5,000 Years*. Brooklyn: Melville House.

Feinberg, Joel. 1970. "The Nature and Value of Rights." *Journal of Value Inquiry,* 4 (4), 243- 260.

Flack, J. C., & De Waal, F. B. M. 2000. "Any Animal Whatever, Darwinian Building Blocks of Morality in Monkeys and Apes." *Journal of Consciousness Studies*, *7* (1), 1–29.

Foot, Phillipa. 1972. "Morality as a System of Hypothetical Imperatives." *The Philosophical Review*, 81:3, 305-316.

Hart, H. L. A. 1961. *The Concept of Law*. Oxford: Clarendon Press.

Kant, Immanuel. 1785/2012. *Groundwork of the Metaphysics of Morals*. Mary Gregor and Jens Timmerman, trans. Cambridge: Cambridge University Press.

Katsafanas, Paul. 2016. *The Nietzschean Self*. Oxford: Oxford University Press.

Leiter, Brian. 2002. *Nietzsche on Morality*. London: Routledge.

May, Simon. 1999. *Nietzsche's Ethics and his War on Morality*. Oxford: Oxford University
Press.

Milgram, Stanley. 1963. "Behavioral Study of Obedience." *Journal of Abnormal and Social Psychology*, 67, 371-78.

Morrisson, Iain. 2018. "Nietzsche on Guilt: Dependency, Debt, and Imperfection." *European Journal of Philosophy*, 26: 974-90.

Owens, David. 2006. "A Simple Theory of Promising." *The Philosophical Review*, 115 (1): 57-77.

Raz, Joseph. 1977. "Promises and Obligations," in P.M.S. Hacker and J. Raz, ed., *Law, Morality, and Society: Essays in Honor of H.L.A. Hart*. Oxford: Oxford University Press: 210–228.

Richardson, John. 2004. *Nietzsche's New Darwinism*. Oxford: Oxford University Press.

Reginster, Bernard. 2017. "What is the Structure of Genealogy of Morality II?" *Inquiry* 61 (1), 1-20.
___. 2011. "The Genealogy of Guilt." In May, Simon ed., *Nietzsche's "On the Genealogy of Morality": A Critical Guide*. Cambridge: Cambridge University Press.

Ridley, Aaron. "Nietzsche's Intentions: What the Sovereign Individual Promises." In *Nietzsche on Autonomy and Freedom*, ed. K. Gemes and S. May. Oxford: Oxford University Press.
___.1998. *Nietzsche's conscience: Six character studies from the Genealogy*. Ithaca: Cornell University Press.

Risse, Matthias. 2001. "The Second Treatise in *On the Genealogy of Morality*: Nietzsche on the Origin of the Bad Conscience." *European Journal of Philosophy*, 9 (1): 55–81.

Rothstein, S. I., & Pierotti, R. 1988. "Distinctions among reciprocal altruism, kin selection, and cooperation and a model for the initial evolution of beneficent behavior." *Ethology and Sociobiology*, 9 (2–4), 189–209.

Strawson, P.F. 1961. "Social Morality and Individual Ideal." *Philosophy* 36: 1-17.

Trivers, Robert. 1971. "The Evolution of Reciprocal Altruism." *The Quarterly Review of Biology,* 46 (1), 35-57.

Watson, Gary. 2009. "Promises, reasons, and normative powers." In Sobel, D. and Wall, S. ed., *Reasons for Action*. Cambridge: Cambridge University Press.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. Cambridge: Harvard University Press.

Chapter 4
Debt, Guilt, and Perverse Guilt

Abstract: In this chapter we continue to uncover GM II's "long history and metamorphosis" (GM II: 3) of conscience. My principle aim is to show that GM II's account of the emergence of bad conscience and guilt corresponds to the development of *moral agency*, and to situate this account in reference to Nietzsche's explanation of the "moralization" of *Schuld* (debt/guilt) at the end of GM II. "Moralized" *Schuld* is defined by two features. It is a form of guilt that attaches to us *as persons*, not simply our actions, and it is *enduring and inexpiable* (it can never be "repaid"). I analyze three kinds of guilt that Nietzsche discusses—Greek guilt, Jewish Guilt, and Christian guilt—to show that "moralization" is not the process whereby material indebtedness becomes guilt, but is rather a perversion of guilt. Finally, I consider whether guilt is inherently irrational and unhealthy on Nietzsche's analysis, and hence whether he thinks it ought to be overcome. I argue that it need not be, and may indeed be healthy and productive on Nietzsche's analysis.

Keywords: debt, bad conscience, guilt, perverse guilt, moral agency

I. Introduction

GM II does for the morality of right and wrong what GM I does for the morality of virtue and goodness: it offers an explanation of how our standards of right and wrong, and the concepts of duty and obligation associated with them, became *moral* standards and concepts, in the same manner the first essay describes how our originally non-moral concepts of virtue and goodness became *moral* concepts.[95] As we saw evidence of in the last chapter, much of GM II is for this reason preoccupied with rules, understanding these to be necessary for the maintenance of society ("primitive requirements of social co-existence" [GM II: 3]), and of our evolving sensitivity to them. This preoccupation with rules compliments Nietzsche's other principle aim in GM II, to provide "the long history of the origins of *responsibility*" (GM II: 2), insofar as GM II also explains the origins of conscience. As we saw in the previous chapter, the conscience was initially

[95] I refer the reader to Maudemarie Clark's influential work on these themes. As she says, "Rules obeyed only out of fear or instinct are not yet perceived as moral rules by those who are disposed to obey them. The main question Nietzsche pursues in *GM* II concerns how such non-moral rules, laws, and customs were transformed into moral ones. His basic answer is that this happened through the development of guilt" (Clark 2015a: 68). See also Clark (1994, 1998).

"bred" in humans through the "mnemo-technique" of punishment (GM II: 3), so that we could remember these rules. The development of conscience therefore made possible a kind of normative competence, an ability to maneuver one's social milieu in ways that allowed us to avoid the suspicion and anger of our peers, by conforming our behavior to expected standards of conduct. The development of non-moral conscience thereby made us "regular" and "predictable" in our behavior (GM II: 2). However, Nietzsche's "long history" of responsibility does not end with our becoming *reliable agents*, understood in these terms. We have yet to discuss the emergence of bad conscience, "the consciousness of guilt" (GM II: 4), as well as the form of conscience that belongs to the sovereign individual, which underwrites his status as "the human being that is permitted to promise" (GM II: 2). Our focus here will be on the development of bad conscience and guilt. (Sovereign conscience will be our focus in Chapter 5.)

It is perhaps most instructive to begin a discussion of these themes with an analysis of GM II's title, "Guilt, Bad Conscience, and Related Matters." The title suggests that Nietzsche's principle aim in the essay is to uncover the origins of bad conscience and the *feeling* of guilt, or *subjective guilt*. To simplify matters, I will refer to subjective guilt as simply "guilt," though unless otherwise specified, by "guilt" I always mean the *feeling* of guilt. Guilt, on Nietzsche's account, is that "reaction of the soul called 'bad conscience, the 'pang of conscience'" (GM II: 14). As this suggests, an episode of bad conscience is the underlying *cause* of guilt feelings, and these feelings manifest as a desire to punish oneself. As we will see, however, Nietzsche's account of guilt is immensely complex. It is not just a feeling, but a feeling attached to an evaluative judgment. Secondly, this judgement can be made with respect to two different objects—my actions or my person (more precisely, one's character). And finally, at least in paradigm cases, guilt involves

two different relations: the relation to *another* whom I've harmed, as well as the relation *I* bear to normative standards I accept for being a good person (see Clark 2001: 58).

In short, guilt is “the feeling of personal obligation" (GM II: 8), as Nietzsche says. Or, as it would probably be clearer to say: guilt is that emotion we experience when we *transgress* norms or obligations we regard as "personal," or that we relate to "personally." As Bernard Reginster puts it, these norms or obligations are “personal" because they "engage our sense of worth as persons" (2011: 66). This means that they are *internalized* standards of conduct, standards that we deem to be legitimate and binding, or as having moral or categorical force, as opposed to simply being prudential or conventional norms ("customs"). As we might say, these norms or obligations are personal because they are constitutive of our practical identity.[96]

My overarching aim in this chapter is to show that moral or responsible agency emerges on Nietzsche's analysis once humans saw themselves as bound to obligations of this form. However, because Nietzsche's account of the emergence of bad conscience and guilt is immensely complex, this broader aim will be pursued within the background of addressing two more pressing interpretive issues. First, Nietzsche argues that guilt has its origins in the material concept of debt that develops within creditor-debtor relationships (GM II: 4, 8). Unlike guilt, being indebted to another does not engage our worth as persons. However, Maudemarie Clark has recently and illuminatingly argued that Nietzsche articulates an account of holding others responsible that rests only on the idea of debt. More precisely, she argues that the account of retributive punishment Nietzsche provides qualifies as a "primitive form of moral address" (2015b: 93). I agree with Clark and endorse her interpretation in Section 2, but her essay does not consider

[96] Thank you to Coleen Macnamara for pressing me to clarify Nietzsche’s expression of guilt as “the feeling of personal obligation” (GM II: 8).

how bad conscience and guilt modify this practice. Relatedly, as we saw in Chapter 2, Nietzsche claims that the practice of retributive punishment does not require regarding the transgressor as a responsible *agent* (GM II: 4). If GM II provides an account of why humans are not exempt from the moral reactive attitudes, as I argued in Chapter 2, the relationship between debt, guilt, and moral agency is in need of further explanation.

I offer this explanation in Sections 3 and 4. My focus in Section 3 is on the *psychological* aspect of guilt, on bad conscience or internalized aggression. Bad conscience originates as a brute desire to inflict punishment on oneself, but it does not become the "consciousness of guilt" (GM II: 4) until these incipient guilt feelings are *conceptualized* or *interpreted* as merited or deserved, as a consequence of failing to honor normative standards one accepts for being a good person. Nietzsche helpfully refers to unconceptualized bad conscience as guilt in its "raw state" (GM III: 20). Though this "raw" guilt is not yet guilt proper, I argue it is important to Nietzsche's account of moral agency because it makes it possible to *hold oneself to expectations*, as well as the depth of soul necessary to do so. Guilt, my focus in Section 4, emerges only once these incipient guilt feelings become attached to a negative evaluative judgment whose object is one's action, one's character, or both. At this point guilt emerges as the "feeling of personal obligation" (GM II: 8).

That brings us to the second, and more complex, interpretive issue. It is not clear when Nietzsche believes that guilt actually emerged. It clearly exists in Christian morality, where guilt takes the form of indebtedness to its "holy God" (GM II: 22). Indeed, this is also when Nietzsche claims that *Schuld* (which can be translated as either "debt" or "guilt") became "moralized." However, Nietzsche describes "moralized" *Schuld* as an enduring and inexpiable state of persons, as a kind of guilt that can never be

repaid. This is very different from what we ordinarily think of as guilt today, which is instead understood as a kind of moral failing that attaches to our actions, experienced by undergoing an *episode* of bad conscience, a "pang of conscience" (GM II: 14). Therefore, one might be skeptical that "moralization" is intended by Nietzsche as an explanation of how material indebtedness *became* subjective guilt. I believe further evidence of this is provided by the fact that much of what Nietzsche says about "moralized" *Schuld* seems to *presuppose* the existence of subjective guilt. If this is correct, we must not only explain the transition whereby material indebtedness becomes guilt (the first interpretive issue), but also how guilt differs from and relates to "moralized" guilt.

I argue in Section 5, along with a small but growing trend in the Nietzsche scholarship,[97] that the *Schuld* that becomes "moralized" is already guilt and not the material concept of debt. In other words, the "feeling of personal obligation" and moral agency existed prior to Christian morality and its "holy God." Specifically, I will argue that the ancient Greeks on Nietzsche's account were capable of and did experience episodes of bad conscience and guilt, and that "moralized" guilt is in fact a *perversion* of subjective guilt, or, more simply, *perverse guilt*. Specifically, I argue that the idea of guilt which Christianity perverts is Jewish guilt, though we have good reason to suspect Nietzsche would have been critical of the latter as well.

In short, the Greek's had a concept of guilt only for *what one has done*; Jewish guilt goes beyond this and manifests also as guilt for *who one is*; and Christian guilt goes beyond the latter by being guilt for who one is *and must be*. Christian guilt is therefore perverse, because it is inexpiable and attaches to us simply in virtue of our being fallible and sinful, and is supported by an unattainable ideal of what humans ought to be (the

[97] See Janaway (2007), Reginster (2011, 2017), and Zamosc (2012).

"holy God"). Consequently, "moralization" is not Nietzsche's explanation of how material indebtedness becomes guilt, but something else. Following Christopher Janaway, I argue that "moralization" is the process whereby being in a state of guilt is judged to be "good," because we ought to always feel guilty for who we are. As Janaway says, "Moralization is the elevation of feeling guilty into a virtue" (2007: 142).

I Section 6 I conclude by way of considering whether guilt is something inherently irrational and unhealthy, something we'd be better off never experiencing on Nietzsche's view. Indeed, in a number of contexts Nietzsche professes to want to free humanity of guilt, to rid "the world" of guilt, and to dispel us of the cruel and inane idea that we deserve to suffer for who we are or what we do.[98] On the basis of these passages, Brian Leiter (2011: 105) has claimed that Nietzsche wants to dispense with the notion of guilt altogether. However, I use the comparative analysis of guilt I offer to show that this conclusion is too overreaching. On my reading, the ancient Greeks had a healthy relationship to bad conscience and guilt, for instance, and so they provide a model we might try and emulate today.

II. Non-Moral Conscience and Indebtedness

We have many metaphors for conscience. It has been called a "social mirror" (Boehm 2012), an "inner voice" (Velleman 1999), and our “moral sense” (Darwin 1871). Nietzsche, to the contrary, conceives of it as a kind of "memory" (GM II: 1), a capacity to overcome forgetfulness. His discussion of conscience in the *Genealogy* moreover invites a narrow and broader interpretation of its function. On the narrower interpretation, the conscience made it possible to remember and sustain the motivation to keep promises or obligations (Reginster 2011, 2017). I criticized this interpretation in the previous

[98] See, e.g., HA 114, 132, and 133; D 78, 87; TI VI: 8; A 25-6.

chapter because it makes promising within contractual relations a condition of the development of conscience, because it develops in response to making promises, when in fact contractual relations assume the existence of conscience as a condition of the ability to make promises, on Nietzsche's view (see GM II: 5). On the broader interpretation I favor, conscience originated in a more primitive social dynamic, one defined by pre-existing social practices it was uniquely responsive to, namely, one in which agents with rank and power routinely held other agents to expectations and punished them for violating them. On this alternative view, conscience instead originated so that we could remain "aware" of those norms and practices, so that we could conform our behavior to them and avoid punishment.

The development of conscience initially made our behavior "regular" and "predictable" for this reason (GM II: 2), but it did not coincide with the emergence of moral agency in humans, as one might naturally expect.[99] It was initially a non-moral faculty, because it was not in any way connected to or responsible for producing moral emotions (the most important one being guilt on Nietzsche's analysis). Conscience initially made possible what Kristin Andrews has called “naïve normativity,” a kind of thinking that she contends is the building block of moral agency.

> [W]hen we look for evidence of specifically moral norms, we lose sight of the basic cognitive requirement for moral agency—namely, ought-thought, which is a cognitive modality much like mental time travel or counterfactual thinking. Thinking about what ought to be the case is—like thinking about what happened in the past, what might happen in the future, and what might be the case under various circumstances—a cognitive mode that requires the thinker to do more than represent what is currently the case. The cognitive mode of thinking about what ought to be the case is what we will refer to here as *naïve normativity*. (Vincent et. al 2018: 58-9)

[99] Darwin and Freud also held that conscience developed in stages and that elements of it were present in non-human animals.

Nietzsche makes similar remarks about the conscience when describing it is a precondition for the ability to make promises:

> In order to have this kind of command over the future in advance, man must first have learned to separate the necessary from the accidental occurrence, to think causally, to see and anticipate what is distant as if it were present, to fix with certainty what is end, what is means thereto, in general to be able to reckon, to calculate—for this, man himself must first of all have become come *calculable, regular, necessary*, in his own image of himself as well, in order to able to vouch for himself *as future*, as one who promises does! (GM II: 1)

These similarities are not coincidental. Naïve normativity requires the agent to represent beyond "what is currently the case," and the conscience is similarly described by Nietzsche as a capacity to overcome "forgetfulness" (GM II: 1), understood to be "an active and in the strictest sense positive faculty of suppression" (GM II: 1). This faculty is "responsible for the fact that whatever we experience, learn, and take into ourselves" remains imperceptible, serving as a "doorkeeper," an "upholder of psychic order, of rest" (GM II: 1). "Forgetfulness," so understood, is actually a kind of *inattention*, an ability to be blissfully unaware. So the conscience, being the opposite of this, is an ability to *sustain* attention, an ability to be remain attentive or aware. Specifically, it is an ability to remain attentive of our obligations, which initially took the form of "I will nots" (GM II: 3). Developing a conscience thus enabled us to think beyond "what is currently the case," and not just by thinking about the past. It also enabled us to anticipate how others will react to us in the future on the basis of how they've reacted in the past.

So, agents who have acquired a conscience are proficient maneuvering social contexts governed by norms, but this does not mean they relate to those as *moral* norms. A rule or norm qualifies as "moral," according to Nietzsche in the *Genealogy*, only if one is held to an expectation to follow it under the presupposition that she should experience guilt for violating it. More precisely, a norm qualifies as moral only if either of the two

following conditions are satisfied: i) its violation *would* elicit guilt in the transgressor, or ii) she is blamed by others under the assumption that she *ought* to feel guilty for violating it. On the account I develop here, this is what it means to hold oneself to a moral expectation, and to be held to a moral expectation by others, respectively.[100] However, in GM II Nietzsche argues that the feeling of guilt developed from the "material" concept of debt that originated in contractual relationships (GM II: 4-8). So in order to understand what is involved in holding oneself or others to moral expectations, we must first analyze doing so under the presupposition of mere indebtedness.

2.2 Debt and Retributive Punishment

Though the conscience originates as a memory of norms or rules, it eventually also encompasses a sensitivity to the *normative standards* that develop within any given community to enforce those rules consistently, justly, and fairly. Once the practice of punishment begins to be administered in accordance with these standards, it becomes *retributive* in nature, i.e., a backward-looking punitive measure that aims to rectify the victim for a past wrong, involving the belief that punishment is deserved. Nietzsche claims that this form of punishment developed in connection to the notion of "equivalence," "the idea that every injury has its *equivalent* in something and can really be paid off" (GM II: 4), which he claims is the "oldest and most naïve moral canon of *justice*" (GM II: 8). The idea of equivalence originated within contractual or "creditor-debtor" relationships, wherein the creditor would punish the debtor as an alternative means of compensation for unpaid debts (GM II: 5). However, as we saw in the previous chapter, initially punishment was not retributive in character; it was simply a response of

[100] See also the account of holding others and oneself morally responsible articulated in Chapter 2, §3.

"anger over an injury suffered" (GM II: 4), which had the effect of creating and enforcing prescriptive rules.

The idea of debt and equivalence modify what is initially a very primitive practice of punishment (indeed, one that I argued all social animals participate in). To see this, we will compare two different forms of punishment Nietzsche describes in GM II: 9 and 10. His focus here, unlike GM II: 5, is on punishment at the communal or societal level, as opposed to punishment at the dyadic level between individuals. Nietzsche claims that the earliest forms of communal punishment were the most severe, and therefore the most memorable:

> The criminal is a debtor who not only fails to pay back the advantages and advances rendered him, but also even lays a hand on his creditor: he therefore not only forfeits all of these goods and advantages from now on, as is fair—he is also now reminded *how much there is to these goods*. The anger of the injured creditor, of the community, gives him back again to the wild and outlawed condition from which he was previously protected: it expels him from itself, and now every kind of hostility may vent itself on him. At this level of civilization "punishment" is simply the copy, the *mimus* of normal behavior towards the hated, disarmed, defeated enemy, who has forfeited not only every right and protection, but also every mercy. (GM II: 9)

Elsewhere Nietzsche explains that the purpose of such "punishment" was to *disgrace* the criminal (HA II.2: 22).

By disgracing him in the same manner as the despised and defeated enemy, exile punishment serves a number of functions. It allows the aggrieved members of community to vent their anger and hostility against the offender. It makes the offender suffer or "pay" for what he did, which gratifies their "instinct for cruelty" (GM II: 5). In being made to "pay" for what he did in a very public way, he becomes a symbolic reminder to the rest of the community of what happens when you break the rules, as well as a reminder of the community's value, as a safe harbor from the untamed and unforgiving state of nature.

Exile punishment does not presume that the criminal *deserves* the suffering inflicted upon him, however. While Nietzsche claims the act of exiling the criminal is a "fair" consequence for having broken the rules (because he reneged on his debt to society), there is no suggestion that what happens to him afterward is anything but a gratuitous act of violence, one subject to no standards of propriety whatsoever. "The creditor could subject the body of the debtor to all manner of ignominy and torture" (GM II: 5), Nietzsche says elsewhere, because by being exiled the debtor was "made an outlaw" (GM II: 10). An outlaw is one who breaks his contract with society and is subsequently banished from it, and so no longer lives "protected, shielded, in peace and trust, free from care with regard to certain injuries and hostilities" (GM II: 9). Nietzsche believes that exile punishment was so excessive and dehumanizing because it reflected the community's insecurity, because the outlaw's act was "dangerous and subversive for the continued existence of the whole" (GM II: 10). Nevertheless, these wanton displays of violence had the indelible effect of reinforcing the paramount importance of following the rules that make the community's existence possible. Exile punishment thus functioned as a *deterrent* to would-be rule-breakers, but it does not aim at redress or repayment for a past wrong, and so is not yet retributive in character.

According to Maudemarie Clark (2015b), retributive punishment comes onto the scene only once the community grows in power and security, at which point Nietzsche says:

> [T]he evildoer is no longer 'made an outlaw' and cast out; the general anger is no longer allowed to vent itself in the same unbridled manner as formerly—rather, from now on, the evildoer is carefully defend against this anger, particularly that of the one's directly injured, and taken under the protection of the whole. Compromise with the anger of the one immediately affected by the misdeed; a striving to localize the case and prevent a further or indeed general participation and unrest; attempts to find equivalents and to settle the entire affair . . . above all the increasingly more resolute will to understand every offense as in some

> sense *capable of being paid off*, hence, at least to a certain extent, to *isolate* the criminal and his deed from each other—these are the traits that are imprinted with increasing clarity onto the further developments of penal law … The "creditor" has always become more humane to the degree that he has become richer. (GM II:10)

Clark argues that at this point punishment qualifies as "a primitive form of moral address" (2015b: 93). Specifically, the practice Nietzsche describes above involves two modifications of exile punishment. First, punishment is a *measured* response that is subject to *normative standards*, in accordance with the principle of equivalence. (It is now possible, for instance, for punishment to be "too severe" or "too lax.") Secondly, criminals are no longer exiled from the community when punished, but are offered an avenue to repair their relationship with the community and remain a member of it *through* punishment. As Clark explains, "The community repudiates the deed, but not the person, to whom it offers a way of restoring his relationship with it. Retributive ideas originally work to isolate the criminal from his deed, so only the deed must be repudiated" (2015b: 95). We will consider both of these modifications in turn.

Consistent with Nietzsche's genealogical approach, the standards creditors used to mete out punishment were initially completely and utterly arbitrary:

> [T]he creditor could subject the body of the debtor to all manner of ignominy and torture, for example cutting as much from [the debtor's body] as appeared commensurate to the magnitude of the debt:— and everywhere and early on there were exact assessments of value developed from this viewpoint—some going horribly into the smallest detail—*legally* established assessments of the individual limbs and areas on the body. I take it already as progress, as proof of a freer, more grandly calculating, *more Roman* conception of the law when the Twelve Tables legislation of Rome decreed it was of no consequence how much or how little the creditors cut off in such a case, "*si plus minusve secuerunt, ne fraude esto*" [If they have secured more or less, let that be no crime]. (GM II: 5)

In other words, it was initially and solely up to the creditor's discretion how much the debtor should suffer or "pay." Retributive punishment was therefore arbitrary, but "from this viewpoint," Nietzsche says, "everywhere and early on exact assessments of value

developed." That is, the idea of equivalence evolved such that these standards were tweaked, codified into law, and applied uniformly.

This does not rule out the possibility that these standards may be arbitrary at the level of the community, "objectively considered" as we might say, but it is important to see that they would not have seemed arbitrary to those raised *within* that environment, who were taught to believe they were "fair" and "just." The Roman equivalences might seem barbaric and arbitrary to us today, but they would not have seemed arbitrary to the *Romans*, especially if, as we are to assume, those standards had been modified over the years to protect the criminal from the direct hostility of the victim, and to instead satisfy the demands of the community at large according to a generally agreed upon equivalence price. (Note, this distinction is still commonly invoked today to differentiate "justice" from "revenge.") The baker whose bread is stolen may want the thief to lose a hand, but the community, capable of pitying the criminal on account of the fact that he was starving, might instead find it sufficient that he lose a finger, or simply pay a fine. In any case, as the community grew in power, punishment became less severe, less gratuitous, and to that extent at least, less arbitrary.

At a certain level of power Nietzsche further maintains that the community began assessing a criminal's action in terms of whether it was "accidental," "intentional," or "negligent" (GM II: 4).[101] Again, there is an "increasingly more resolute will to understand every offense as in some sense *capable of being paid off*, hence, at least to a certain extent, to *isolate* the criminal and his deed from each other" (GM II: 10). Prior to taking these distinctions into account, we can imagine that the criminal was punished regardless of whether he *intended* to break the law, regardless of whether he was

[101] He even entertains the possibility that the strongest communities could do away with punishment altogether (GM II: 10).

ignorant of the law, indeed regardless of mitigating circumstances entirely. In Strawsonian terms, punishment was not yet sensitive to the issue of *excuses*, the question of whether one acted with an ill or indifferent will such that their behavior would warrant resentment on the part of others. However, with increasing efforts to isolate the criminal from his deed, these sorts of considerations became integral to retributive punishment.[102]

It should be clear that the form of punishment we are now considering is no longer a mere venting of anger or hostility that deters would-be offenders from violating rules. Punishment is meted out according to the principle of equivalence, informed by normative standards agreed upon by the community, and the reactive affects mollified according to whether the action was accidental, intentional, or negligent. Instead of being forward-looking, punishment is now a backward-looking measure that seeks to rectify a past injustice. It is now something the criminal is believed to "owe" as a matter of "fairness" or "justice" (GM II: 8). Finally, by rectifying that injustice, offenders are offered a means of restoring their relationship with the community via punishment.[103]

---

[102] According to classical retributive theory, as it is understood today, a criminal deserves to suffer only if two conditions are satisfied. A person must not only perform a guilty action (*actus rea*), she must also do so with guilty intentions (*mens rea*). From the Latin *'Actus non facit reum nisi mens sit rea'* (an act does not make a person guilty unless the mind is also guilty). That is, she must not only break the law, she must also have knowledge of the law and act with the understanding that she is breaking it.

[103] This fact is particularly important to Clark's Strawsonian reading of GM II because it implies that we no longer adopted the "objective attitude" toward offenders. As Strawson notes in his discussion of punishment, "The holding of these does not, as the holding of objective attitudes does, involve as apart of itself viewing their object other than as a member of the moral community. The partial withdrawal of good will which *these* attitudes entail, the modification *they* entail of the general demand that another should, if possible, be spared suffering, is, rather, the consequence of *continuing* to view [the offender] as a member of the moral community; only as one who has offended against its demands" (1962: 90).

I am sympathetic to Clark's interpretation, but matters are complicated here because Nietzsche's "reactive affects" are not coextensive with Strawson's "reactive attitudes." In particular, on Nietzsche's account expressing affects like anger or revenge toward another agent doesn't presume she is *morally* responsible. This discrepancy is explainable by the fact that Strawson analyzes our responsibility practices as they've evolved and exist now, whereas

Crucially, this practice can also make sense of the kind of resentment or moral anger we experience in response to cheaters and acts perceived to be unfair (see Wallace [1994], Vincent et. al [2018], Walker [2006]). This, too, represents an important departure from the way anger operated in exile punishment. As we saw in Chapter 2, anger is a form of other-directed blame in response to an "injury suffered" (GM II: 4), and it takes as its object the perpetrator of the injury, expressed as a desire to punish him for the wrong committed. It is the most basic of the "reactive affects" Nietzsche discusses in GM II. As described in the previous chapter, anger is elicited when someone violates *expectations of conformity*, expectations to follow rules or conventions, and is prototypically enforced by those with power. What I am here calling "resentment" or "moral anger," by contrast, additionally rests on considerations of what is fair, legitimate, and deserved, on what we might call *expectations of redress*, an expectation that wrongdoers deserve to suffer as "payment" in proportion to their debts.

This practice therefore assumes a different *comportment* to rules, and for this reason I think Clark is right to characterize Nietzsche's account of retributive punishment as a "primitive form of moral address." Here we have a practice of holding responsible—of holding others to expectations along with the proneness to experience reactive affects—that is sensitive and responsive to widely accepted normative standards, and to ideas of fairness, legitimacy, and desert, not mere conformity with rules. That is, we now have a practice of holding responsible whereby we do not punish and blame others simply because we are *disposed* to do so, but because doing so is *appropriate* or *fitting*. As Clark says, "With the belief that the infliction of pain and suffering is justified,

---

Nietzsche provides a genealogy of their development. Of particular importance here is that Nietzsche seems to assume debtors at this stage were incapable of feeling guilt. We can still refer to this as a "moral community," so long as it is acknowledged that these individuals are not yet responsible or moral agents. Perhaps we might call it a "primitive" moral community, following Clark's convention.

that the offender owes or deserves it, it seems that we have a primitive case of holding responsible or accountable and of blaming behavior, and one that involves no incompatibilist assumptions" (2015b: 95).

However, if bad conscience is not yet in the picture, as Clark seems to acknowledge in other papers, the idea of "guilt" at this stage would be purely external, what I have called objective guilt, and which she calls "primitive guilt" (see Clark 2001: 57, 58; 2015a: 69). What makes this idea of guilt primitive is that it implicates only the "material" (GM II: 4) concept of debt that Nietzsche argues is a precursor to subjective guilt, which I argue below is a condition of responsible agency. Indeed, as we saw in Chapter 2, Nietzsche acknowledges as much himself in the fourth aphorism. He claims that "throughout the greatest part of human history [retributive] punishment was definitely *not* imposed *because* one held the evil-doer responsible for his deed, that is, *not* under the presupposition that only the guilty one is to be punished" (GM II: 4). Rather, these "evil-doers" were punished as a result of "anger over an injury suffered," which was "held within bounds" and "modified" by the idea of equivalence (GM II: 4). The reason for this is that the evil-doer was not blamed under the presupposition that he can or ought to *feel* guilty for violating rules or norms. Nietzsche says this is analogous to the way we punish children today.

We are therefore to imagine punishment being administered to agents who are incapable of feelings of guilt, remorse, or contrition because they have not yet acquired a bad conscience (see GM II: 14-15). They are agents who never feel the need to apologize, and so are also never in need of forgiveness. Incapable of remorse or contrition, the criminal can offer no sincere apology and can never be in the kind of state for which forgiveness might provide him some relief. More importantly, because he is incapable of

guilt, the "feeling of personal obligation" (GM II: 8), he cannot *hold himself* responsible or accountable for his action. We might say that he is like a psychopath for whom moral rules are viewed as mere conventions (though perhaps he does care about whether those rules are followed consistently), a fact which cannot but severely impair his relationship with others. This practice lacks the "depth" characteristic of *moral* blame and responsibility, we might say, because it functions solely to satisfy the demands of the community, whereas when it comes to the offender, he makes no similar demands of *himself*.

As Clark observes, "It is only when one internalizes (i.e., adopts against oneself) the hostile attitude of one who thinks you owe him something, and, more specifically, that you deserve to suffer for what you owe him, that it starts to be recognizable as guilt" (2015a: 70).[104] The criminal at this stage in Nietzsche's story is certainly capable of recognizing his indebtedness, his objective guilt, his being in a state where he is judged guilty by others. However, incapable of having an episode of bad conscience, he would not feel that he *deserves* punishment, nor would he take this upon himself, and so in this sense could not take responsibility or accountability for his action. To be sure, he might assent to the proposition that he "owes" some measure of punishment for violating the rules, as custom dictates, but he would have no "consciousness of guilt" (GM II: 4) in doing so, i.e., no accompanying *guilt feelings*.[105]

In the following two sections I attempt to show how the development of bad conscience allows for the transformation Clark alludes to above, namely, the

[104] Similarly, as Christopher Janaway says, "The feeling of guilt is a process whereby some putatively permitted or rightful punishment is exacted internally by means of a partial identification with those whom one conceives as angered by one's transgression" (2007: 136).

[105] As Nietzsche claims, the criminal would confront his punishment with "'sadness, accompanied by the image of a past matter that has turned out in a manner contrary to all expectation'" (GM II: 15), but he would have "had no other 'inner pain'" (GM II: 14).

internalization of objective guilt, whereby the state of being judged guilty by others becomes the state of judging *oneself* guilty. Put another way: bad conscience allows for the internalization of the idea of debt, making it subjective guilt. In Section 3 I focus on the development of bad conscience as the underlying cause of incipient guilt feelings, with an eye toward explaining how it creates the conditions necessary for moral agency. In Section 4 I then focus on Nietzsche's account of how those incipient guilt feelings come to be *conceptualized* or *interpreted* as a debt one incurs for falling short of standards of personal worth, thereby becoming guilt.

III. Internalization and Holding Oneself to Expectations

In contrast to the psychopath we just imagined, moral agents have a different comportment to moral norms. As an initial proximation, we might say that they are motivated to comply with them *as such*, in recognition of their being categorical and having a kind of authority. As we saw in Chapter 3, Nietzsche initially tried to explain this comportment to moral norms in terms of "the motive of obedience to tradition" in *Daybreak* 9, where he offered the hypothesis that it was little more than a habit inculcated in us through the morality of custom, a consequence of an inherited "instinct of obedience" (BGE 199). I characterized these as *non-hypothetical imperatives* in Chapter 3, because although we are strongly inclined to follow the commands of those in power, this does not presume that we view those commands as having authority, as categorical, or as legitimate and binding. In GM Nietzsche's view of moral motivation changes: norms are taken to be *moral* in virtue of the fact that violating them elicits guilt. A consequence of this is that we see ourselves as legitimately bound to them, because they are personal obligations that engage our sense of worth as persons.

The connection to guilt points to another important aspect of moral agency: when moral agents transgress these norms, they believe doing so is objectionable and they *hold themselves* accountable, because these norms make a claim on them regardless of whether *others* would blame or punish them. As I argued in Chapter 2, for agents capable of experiencing guilt, "ought" for them has a different significance than "ought" does for a child. Specifically, they are capable of both the retrospective and prospective form of the ought-judgment "I should not do/should not have done that because it is/was *morally* wrong" (see GM II: 14-15). Consistent with the developmental analysis of conscience that has been offered thus far, this ability was not innate in human beings. Rather, it's the emergence of *bad* conscience that makes this unique comportment to norms possible. This is because, to make this judgment, an agent must also experience feelings of remorse or contrition; in short, she must be undergo "that reaction of the soul called 'bad conscience,' 'pang of conscience'" (GM II: 14), and this feeling must be conceptualized as the failure to honor normative standards one accepts for a good person. Here I will focus on the first part of this relation, the *psychological* aspect of guilt, the production of these inward painful feelings. In the next section we will focus on Nietzsche's explanation of the process whereby these feelings come to be *interpreted* as the consequence of failing to honor personal obligations.

3.1 Internalization (guilt in its "raw state")

Nietzsche's explanation of the origins of bad conscience is complicated in part because he speaks of bad conscience in two different senses. He initially refers to bad conscience in the same manner that he speaks of the conscience more generally, as a kind of consciousness or awareness, the "consciousness of guilt" (GM II: 4). However, apart from this one reference, he far more often refers to bad conscience as the psychological

mechanism of internalized aggression, which is the *cause* of these guilt feelings we become "aware of." What distinguishes bad-conscience-the-mechanism from bad-conscience-as-guilt-awareness is that the latter is a *conceptualized* or *interpreted* form of incipient guilt feelings already present in the first form. Following Nietzsche's lead, when I refer to bad conscience I will generally mean the *mechanism* that produces these uninterpreted guilt feelings. I will denote this as bad conscience$_m$ so as to not invite confusion. These feelings have their origin in a process Nietzsche calls "internalization."

> All instincts that do not discharge themselves outwardly *turn themselves inwards*—this is what I call the *internalizing* of man … Those terrible bulwarks with which the organization of the state protects itself against the old instincts of freedom—punishments belong above all else to these bulwarks—brought it about that all those instincts of the wild free roaming human turned themselves backwards against man himself. Hostility, cruelty, pleasure in persecution, in assault, in change, in destruction—all of that turning itself against the possessors of such instincts: *that* is the origin of "bad conscience." (GM II: 16)

Elsewhere Nietzsche refers to this as guilt in its "raw state" (GM III: 20), a convention that I will follow here.[106] Guilt here is "raw," as it were, because the feelings that bad conscience$_m$ produces have not yet been conceptualized as a moral failure, nor are they the *result* of a moral failing of some kind. It is just an inner pain the agent experiences and inflicts upon himself, as a result of internalization. Our focus in this section will be on "raw guilt."

Above Nietzsche suggests that internalization is a consequence of socialization, of human beings coalescing into groups for mutual protection and imposing rules on one another. However, if this were his view, there would be no difference in kind between

[106] At the beginning of this aphorism he admits to only offering "preliminary expression" of the phenomenon of bad conscience (GM II: 16), and later, "as a piece of animal psychology, no more" (GM III: 20). Some scholars call this "animal bad conscience" to distinguish it from bad conscience *qua* "the consciousness of guilt" (Leiter [2015], Risse [2001]). Others make a similar distinction between bad conscience and "guilt consciousness." The terminology is immaterial; everyone agrees that internalized aggression alone is not sufficient for guilt, which requires conceptualization.

conscience as a memory of "I will nots" (GM II: 3) and "bad" conscience—the former would only be an instance of the latter.[107] Fortunately, Nietzsche qualifies the above remarks in the next aphorism, where he tells us that internalization is actually a result of *subjugation*; it occurs when "a race of conquerors and lords, which, organized in a warlike manner … unhesitatingly lays its terrible paws on a population enormously superior in numbers perhaps, but still formless, still roaming about" (GM II: 17). This passage instead suggests that bad conscience results when rival groups impose novel and stringent laws on *other* groups, not whenever groups impose rules on themselves. This dynamic creates sudden and systemic conditions of oppression that we were ill-equipped to manage as "half animals … happily adapted to wilderness, war, roaming about, adventure" (GM II: 16), which necessitated the internalization of the instincts that were adaptive to that form of life and could no longer be discharged, our aggressive drives in particular. What resulted was a novel kind of *internal pain* manifesting as a desire to inflict punishment on oneself, a "will to self-maltreatment" (GM II: 18) or "cruelty turned backwards" (GM III: 20). These are feelings of "raw guilt."

Before proceeding to an explanation of how bad conscience$_m$ creates the psychological conditions necessary for moral agency, I must first make some general remarks about the nature and prevalence of bad conscience$_m$ in humans to prevent confusion. First, I take it that bad conscience$_m$ is a *pervasive* phenomenon in human beings, evidenced by the fact that feelings of remorse and contrition are quite common among humans. It is rarely the case, though certainly possible, that we come across someone who is incapable of experiencing these emotions. The explanation for the

[107] Aaron Ridley (1998) thinks Nietzsche's views on conscience are contradictory for this reason. Alas, delving too deeply into these matters would lead us too far afield. Suffice it to say, I disagree with Ridley. I refer the reader to two recent articles by Iain Morrisson (2018: 986, n.7; m.s.) that explore the anthropological record Nietzsche was relying on. I am sympathetic to most of his conclusions.

pervasiveness of bad conscience$_m$, if Nietzsche is right, is that the conditions he describes above (i.e. those in which tribes conquer other tribes) were prevalent among pre-historic humans. Secondly, bad conscience$_m$ is *realized in degrees*. Again, this would seem to accord with common experience. Though we rarely meet someone incapable of feeling guilty, remorseful, or contrite, we often do meet people who are *more* prone to these feelings than others. If Nietzsche is right, this would be explainable in part by the fact that such people have inherited stronger and more active dispositions to punish themselves, because they've descended from ancestors that were subjugated more severely and/or for longer durations.[108]

These two observations are important because, if they are correct, bad conscience is not a uniquely *Christian* phenomenon, and that has profound implications for how we ought to interpret GM II's genealogy of guilt. As a conceptual matter, one implication is that we do not have to interpret bad conscience as an enduring and inexpiable state of persons, as Christianity interprets guilt on Nietzsche's analysis. Bad conscience might also and instead be interpreted as the cause of *episodes* of guilt. Moving forward I will be taking it for granted that bad conscience$_m$ is pervasive and realized in degrees, and hence not a uniquely Christian phenomenon. (We will see evidence in support of this supposition as we proceed.)[109]

How, exactly, does "raw guilt" compare to guilt proper? To explain this, I think it will be helpful to invoke Brian Leiter's (2013) distinction between "basic affects" and "meta-affects." Raw guilt is a "basic affect" or feeling on Nietzsche's account. It is a mental state that involves an "inclination for" or "aversion to" certain actions (see D 34),

[108] Obviously, culture and upbringing play role here as well (arguably a larger one), but the reason a particular culture exhibits a greater propensity for and receptivity to guilt would in turn be explained by the fact that it descended from tribes that were more severely subjugated and/or for longer durations.

[109] Thanks to Maudemarie Clark, who pressed me to be clearer on these issues.

in this case an inclination to punish oneself, and it has a distinct qualitative feel, i.e., the kind of internal pain we associate with remorse. However, guilt proper is what Leiter calls a "meta-affect," which is an affect conditioned by an *evaluative judgment* we make about some basic affect. As Nietzsche observes in *Daybreak*, "The same drive evolves into the painful feeling of *cowardice* under the impress of the reproach custom has imposed upon this drive: or into the pleasant feeling of *humility* if it happens that a custom such as the Christian has taken it to its heart and called it *good*" (D 38). Leiter speculates that this "drive" is something like the disposition to avoid offending one's powerful enemies. Nietzsche's point is that it is felt differently by, say, the Greek and the Christian, depending on the judgment associated with it—whether it is deemed "good" or "bad" to avoid such offense. To be guilt proper, bad conscience$_m$ must likewise be attached to an evaluative judgment concerning the *appropriateness* of these incipient guilt feelings. That is, one must judge raw guilt to be "good" or "bad," to be merited or deserved, in relation to normative standards one accepts for being a good person. As we will see, this judgment can take as its object the agent's action, the agent himself, or both.

Having distinguished bad conscience$_m$, the "basic affect" of raw guilt, from guilt proper (the "meta-affect"), we will now consider how bad conscience$_m$ creates two conditions necessary for moral agency.

3.2 The Development of Soul

Nietzsche claims that bad conscience$_m$ gave "depth," "breadth," and "height" to the human soul (GM II: 16). The significance of this is that bad conscience$_m$ transformed us from being healthy but uninteresting *animals*, to being sick but interesting *human beings*. To see how it did so, let's begin by comparing raw guilt to anger. Note, raw guilt is just internalized aggression manifesting as self-directed punishment, and recall that

punishment on Nietzsche's account is a response of "anger over an injury suffered" (GM II: 4). When anger manifests as punishment of another, the agent's urge to attack a negative stimulus, the cause of the injury, is given expression and thereby discharged. Punishing another is thus a way of managing two different sources of discomfort. First, it manages the pain of the *injury*, by deterring the perpetrator from initiating further injury. Secondly, it allows for the *expression* of anger and the negative affects associated with it, which induce the agent to lash out at the perpetrator. By punishing the source of the injury, the subject thereby redirects the injury's cause and succeeds in discharging the negative affects produced by experiencing anger.

In the kind of conditions of oppression Nietzsche is describing, this natural outlet for the release of negative affects is blocked. (That is what internalization *is*, after all—a technique for managing one's aggressive impulses by redirecting them toward the self.) A negative feedback loop thus results: the agent is able to express his animal urge to inflict suffering by redirecting them toward the self, but since he is now their object, his doing so results in further injury to himself, resulting again in the desire to inflict suffering. He becomes "a soul compliant-conflicted with itself, that makes itself suffer out of pleasure in making-suffer" (GM II: 18), "an animal soul turned against itself, taking sides against itself" (GM II: 16). This state of inner conflict is crucial to our becoming human beings; it signaled our "forceful separation from [our] animal past" (GM II: 16).

Note, this last quote also serves as evidence of the pervasiveness of bad $\text{conscience}_{\text{m}}$. If bad $\text{conscience}_{\text{m}}$ is meant to signal this "forceful separation," it must be something that humanity in general has undergone. This means that internalized aggression and the negative feedback loop it creates is also something that most of

humanity has endured at some point. I will contend in Section 4 that the Greeks employed an ingenious solution to this problem. They placed a limit on bad conscience$_m$, by essentially redirecting and projecting part of their guilt onto their gods. For now let's imagine, as Nietzsche does in GM II: 16-22, that bad conscience$_m$ is *not* redirected.

To the *extent* that our aggressive impulses become internalized, so too does the "soul" acquire depth:

> All instincts that do not discharge themselves outwardly *turn themselves inwards*—this is what I call the *internalizing* of man: thus first grows in man that which he later calls his "soul." The entire inner world, originally thin as if inserted between two skins, has spread and unfolded, has taken on depth, breadth, height to the same extent that man's outward discharging has been *obstructed*. The man who, for lack of external enemies and resistance, and wedged into an oppressive narrowness and regularity of custom, impatiently tore apart, persecuted, gnawed at, stirred up, maltreated himself; this animal that one wants to "tame" and that beats itself raw on the bars of its cage; this deprived one ... had to create out of himself an adventure, a place of torture, an uncertain and dangerous wilderness—this fool, this longing and desperate prisoner became the inventor of "bad conscience." (GM II: 16)

Note, Nietzsche says that the soul takes on "depth," "breadth," and "height ... to the same extent" that a person's ability to discharge his aggressive instincts outwardly is "obstructed." So, I take it that what he is describing here is the general process of internalization, which is realized in degrees, though he is doing so while foreshadowing a discussion of specifically *Christian* guilt. (As we will see in Section 5, this is the most extreme example of internalization.)

More generally, one might wonder what it means for the soul to acquire this "depth," "breadth," and "height." We know this occurs as a result of internalization, but how? This process essentially involves replacing those "external enemies" the agent can no longer vent his hostility against with *internal* enemies—the agent's psychological drives. These internal enemies are created through a "declaration of war against the old instincts," our "old leaders, the regulating drives that unconsciously guided [us] safely"

(GM II: 16). As Nietzsche says, "all at once all of [our] old instincts were devalued and 'disconnected'" (GM II: 16). To illustrate this, let's consider an example.

Imagine that you have offended me in some way. On Nietzsche's analysis, my first reaction would have been to lash out back at you, to harm you for the offense you caused, to get revenge. Under my previous customary regime, I was allowed to express this desire "within bounds," as mediated by the principle of equivalence, or on others who resided outside the community. But now that I have been subjugated I have no recourse to revenge in these ways. Lacking an external outlet, Nietzsche suggests that the object of my anger now instead becomes *my desire for revenge itself*. I have subsequently developed a second-order desire to *refrain* from acting on my first-order desire to seek revenge. The soul acquires "depth" as we develop higher-order desires in this way. And not just desires, but eventually values and practical commitments, which can then come into conflict with our drives and first-order desires.[110] This state of inner conflict characterizes our transformation from being mere "animals" to becoming full-fledged human beings.[111]

3.3 Holding Oneself to Expectations

Also essential to this transformation, from the point of view of moral agency, is the fact that bad conscience$_m$ makes it possible to *hold oneself to expectations*. To see this, let's contrast the agent who experiences raw guilt with one who has not yet internalized her aggressive instincts. This latter agent has what we might call a *customary identity*, but this is comprised solely of the remembrance of customs and debts. Similarly, although she possesses a conscience, this is only an awareness of the norms and expectations of

[110] What is the source of these eventual values? I take it Nietzsche's answer is culture and morality grounded in religion (see §4.1 below).

[111] This account is similar to the one offered by Clark and Dudrick (2012), see especially pp.200-210.

others. She is therefore "conscious of" the *disapproval of others* (what Nietzsche calls "objective awareness" [D 26], being the object of another's gaze, which he claims all social animals are capable of), and so she could feel *embarrassment*, and perhaps even shame.[112] However, she could not *disapprove of herself*, because her aggressive impulses had not yet been internalized.

Raw guilt manifests as a desire to punish oneself, and subsequent to our souls becoming "deep," this basic affect now manifest as a desire to punish oneself for violating expectations we hold ourselves to. As above, I now desire to punish myself simply for being angry or wanting revenge, which presumes that I hold myself to an expectation to *not* feel angry or vengeful, or that I have made an "enemy" of my natural tendency to feel angry or vengeful. When we violate these expectations, bad conscience$_m$ is engaged and once again we experience incipient feelings of guilt, creating that negative feedback loop just discussed. If the agent doesn't find a way out of this cycle of self-inflicted violence, bad conscience$_m$ indeed becomes a "deep sickness" (GM II: 16), as Nietzsche describes it. It becomes a pervasive and enduring state of one's person, as opposed to something she experiences only occasionally.

Note, however, that these incipient feelings have not yet been *interpreted* as the deserved consequence of failing to honor norms or standards one accepts for being a good person. This creature, as we are imagining him, is not yet judging his anger to be "bad," "evil," or "sinful." He has developed only a second-order desire to avoid acting out

[112] On David Velleman's compelling analysis, shame is an emotion of exposure that involves the awareness of one's "threatened loss of social standing" as "a self-presenting creature" (2003: 37). It is responsive to a kind of *anxiety* about that threatened loss of social standing, which has its roots in the need to be able to cooperate with and rely on others. Consistent with this, we can imagine that pre-historic humans during the morality of custom were constantly in the business of advertising what Michael Tomasello calls a "group" or "cooperative" identity, "to make sure that others could identify them as in-group members, to coordinate with the group, and to avoid punishment" (2016: 100).

of anger or vengefulness, and so he does not yet experience guilt when he lapses. He possesses only the "basic affect" of raw guilt, not the "meta-affect" of guilt.

There is no great gulf between him and the psychopath we began our discussion considering; that difference consists solely in his experiencing a felt need or desire to *punish himself* for violating these expectations. This is what I mean by saying he *holds himself to an expectation* (when he violates these expectations, he experiences raw guilt). So we might say that he, unlike the psychopath, is *self-monitoring*.[113] However, holding himself to expectations in this way is little more than a coping mechanism, necessitated by his radical and swift change in environment. He is, we might say, *primed* to accept judgements concerning his guilt, but he does not yet judge himself to be guilty. He is "Man, suffering from himself in some way or another, physiologically in any case, somewhat like an animal locked in a cage, uncertain why, to what end? Desirous of reasons—reasons alleviate …" (GM III: 20).

IV. Guilt as "the Feeling of Personal Obligation"

We have uncovered the psychological mechanism that is the cause of incipient guilt feelings, bad conscience$_m$, and the "basic affect" of raw guilt that it produces. We now turn to uncovering the origins of guilt, or bad-conscience-as-guilt awareness. Because Nietzsche characterizes guilt as “the feeling of personal obligation” (GM II: 8), that is how I will refer to guilt to distinguish it from bad conscience$_m$.

To experience guilt, the agent must interpret or conceptualize the incipient guilt feelings produced by bad conscience$_m$ in a particular way. As Christopher Janaway explains:

> What differentiates the feeling of guilt from other kinds of psychological pain? It must be the way the subject represents herself: she must at least take herself to

[113] (I’m not actually sure whether psychopaths are *not* self-monitoring in this way.)

> have done harm, to have transgressed, usually against some other agent, in such a way as to violate an obligation she accepts herself to be under. To feel guilty requires an inner suffering that one represents as undergone because one has departed from what one believes one ought to do, in a way that is likely to cause anger or resentment from others and would permit them to despise or maltreat one. (2007: 136)

Clark makes a similar point when she notes that guilt involves "two different relations: the relation of the individual to someone she has injured or failed in some way, and then the self's relation to the standards she accepts for a good person" (2001: 58).

According to Clark and Janaway, guilt is not equivalent to bad conscience$_m$ because it involves at least three additional things:

(i) making a judgment about myself or my behavior,
(ii) in reference to obligations or normative standards I accept for being a good person,
(iii) and so this judgment engages my sense of self-worth, i.e., is "personal."

Because guilt is a manifestation of bad conscience$_m$, we must add a fourth and final condition:

(iv) I feel and judge myself to be deserving of punishment for violating obligations of this form, i.e., "personal obligations."[114]

In short, guilt is a feeling of deserved self-punishment for violating personal or *moral* obligations. Why are personal and moral obligations equivalent on Nietzsche's analysis? Because obligations become moral by becoming attached to guilt, and obligations become attached to guilt by becoming "personal," by engaging our sense of worth, and so personal and moral obligations amount to the same thing. We might also

[114] I do not mean to imply that we must always make these judgments consciously and explicitly, though that is certainly the form of guilt Nietzsche ultimately wants to explain. My understanding of Nietzschean guilt has been influenced by P.S. Greenspan (1992), who argues for a "non-judgementalist" account guilt. According to her, guilt "amounts to discomfort with a certain evaluative propositional object and hence may be said to correspond to a judgment," namely, "the subjectively guilty agent feels *as if* he were *morally* responsible" (1992: 287). The "*as if*" qualifier is important because she argues (plausibly, in my opinion) that guilt feelings do not arise solely from making this judgment, explicit or otherwise. This allows her view to accommodate perverse and unconscious forms of guilt—a project that is very much consistent with Nietzsche's, as we can see.

say that guilt is the *self-reactive feeling of responsibility*, for these reasons, because it is the expression of one who takes *responsibility or accountability* for violating personal obligations. Finally, Nietzschean guilt, at least on my analysis, can take as its object two different things: the agent's action, the agent's conception of himself as a person (his character), or both. As we will see below, recognizing this is essential to understanding when and why guilt is unhealthy, perverse, and objectionable on Nietzsche's view.

There is first a problem that must be addressed. The problem is that Nietzsche nowhere in GM II gives a clear explanation of ii), of how and when bad conscience$_m$ becomes attached to judgments corresponding to standards I accept for being a good person. As Janaway notes, "Something Nietzsche does not explicitly provide for in his analysis—but which must be there nevertheless for guilt to occur—is the conception of oneself as a transgressor in one's own eyes" (2007: 136). If Janaway is right, Nietzsche does not provide a full explanation of the process whereby material indebtedness becomes the state of subjective guilt. Naturally, this would seem to correspond to what Nietzsche calls the "moralization" (GM II: 21) of *Schuld* and *Pflicht*. *Pflicht* here means "duty," but *Schuld* in this context could mean either "debt" or "guilt." However, Janaway and I agree that a number of things that Nietzsche says about "moralization" complicate this natural reading, and so we think the *Schuld* which becomes "moralized" is in fact already guilt, not just debt.[115] That is what I will argue below.

The first problem regarding "moralization" concerns the way in which Nietzsche describes it. Briefly, "moralized" guilt is indebtedness to Christianity's "holy God," and as I take it to be clear from the description below, this is a rather extreme manifestation

[115] Clark and many others (e.g., Risse [2001], Leiter [2015]) interpret "moralization" as the process in which material debt becomes subjective guilt. This is indeed the standard view. On this interpretation, debt becomes guilt through Judeo-Christian morality. Janaway and I disagree with them about this.

of guilt and not necessary to simply experience an *episode* of guilt. Moreover, it actually seems to *assume* the existence of episodic guilt:

> This is a kind of madness of the will in psychic cruelty that has absolutely no equal: the *will* of man to find himself guilty and reprehensible to the point that it cannot be atoned for; his *will* to find himself punished without the possibility of the punishment ever becoming equivalent to the guilt; his *will* to infect and make poisonous the deepest ground of things with the problem of punishment and guilt in order to cut off the way out of this labyrinth of "*idées fixes*" once and for all; his *will* to erect an ideal—that of the "holy God"—in order, in the fact of the same, to be tangibly certain of his absolute unworthiness. (GM II: 22)

Note that the "will" to be explained is variously described as the "will to find oneself guilty and reprehensible," to find oneself "punished without the possibility of punishment ever becoming equivalent to the guilt," to "infect" oneself, others, and the world with "the problem of punishment and guilt," and all of this requires "erect[ing] an ideal—that of the 'holy God'," to explain why we are guilty in these extreme ways.

Why do these formulations *presume* the existence of guilt? Because all of them only make sense if one already has a conception of oneself as failing to live up to normative standards of some kind. They assume Clark's second relation, the self's relation to standards she accepts for a good person, is already in place. Specifically, that relation bears some important connection to "holiness," the absence of sin, and this assumes that one who is sinful or in a state of sin is "bad," "evil," or "lesser," not that one is simply experiencing bad conscience$_m$. I will expand on these remarks in Section 5, where I argue that Christian guilt is a perverse extension of Jewish guilt, i.e. it co-opts Judaism's holiness morality and its standards, at the center of which is the notion of sin.

Note also that these "moralized" expressions of guilt are meant to explain not being guilty *for what one has done*—that is, *episodes* of bad conscience$_m$—but being guilty simply *for being who one is*—or bad conscience$_m$ as an *enduring state of the person*. This discrepancy surrounding guilt *for what one has done* and guilt *for who one*

is arises from the fact that, as we observed in the previous section, bad conscience$_m$ is realized in *degrees*, according to the strength and/or duration that a people have been subjugated, and consequently, their aggressive instincts internalized. Therefore, what Nietzsche describes above as "the will in psychic cruelty that has absolutely no equal" is how bad conscience$_m$ developed within Christianity, connected to a perverse idea of guilt.

This is *not* his account of how bad conscience$_m$ manifests in other religions, notably ancient Jewish and Greek religion, the latter of which will be my focus in Section 4.2. According to Nietzsche, Greek religion proves that there are "*more noble* ways of making use of the fabrication of gods than for this self-crucifixion and self-defilement of man" (GM II: 23).

4.1 The Origin of Gods and Subjective Guilt

According to Janaway, the second relation of guilt is implicit (but unexplained) in the aphorism prior to where Nietzsche announces guilt's "moralization." The reason for this is that Nietzsche begins speaking of "guilt feelings" [*Schuldgefühl*], and argues that these originated long ago within the context of ancestor worship. The creditor-debtor relationship, he explains, came to be interpreted as a "relationship of *those presently living to their ancestors*" (GM II: 19). Early tribal communities began believing that they existed and prospered owing to the favor of these ancestors as "powerful spirits," and so they continued to follow the customs believed to be their "statutes and commands" (GM II: 19). Importantly, the community also offered them sacrifices so that they would not withdraw their favor. All of this occurred within primitive cultures where fear, anxiety, and superstition were pervasive.[116] Nietzsche further speculates this is how

[116] "Within the original clan association—we are speaking of primeval times—the living generation always acknowledges a juridical obligation to the earlier generation, and particularly the earliest one, which founded the clan … Here the conviction holds sway that it is only through

belief in gods originated, "an origin, that is, out of *fear*!" (GM II: 19). Finally, as these clans grew in power, so did their gods, leading eventually to monotheism.

So far so good. The community felt *indebted* to the founding members of the tribe, who became their gods, and they abided their commands, which when passed down through the generations became *customs*, or "traditional ways of behaving" (D 9), and they worshiped and offered sacrifices to these gods out of fear. However, Nietzsche then makes a very puzzling claim: "For several millennia the feeling of guilt [*Schuldgefühl*] toward the deity did not stop growing and indeed grew ever onward in the same proportion as the concept of god and the feeling for god grew on earth and was borne up on high" (GM II: 20). He then goes on to claim that the Christian god, being the "maximum god that has been attained thus far," has also "brought about a maximum of the feelings of guilt" (GM II: 20). On the one hand this is consistent with the idea that Christian guilt is a perversion of guilt as it already exists, but on the other hand it is deeply inconsistent with one of the main threads of the second essay as a whole.

Nietzsche argues in no uncertain terms that guilt has its provenance in bad $\text{conscience}_{\text{m}}$, not fear of punishment (GM II: 14-15). So it cannot be, per the above suggestion, that greater fear of punishment, owing to our greater indebtedness to more powerful gods, can ever produce feelings of guilt. If *human* creditors cannot produce feelings of guilt in offenders simply by punishing them, *divine* creditors cannot produce feelings of guilt simply by threatening to punish them, either.[117] Janaway believes that

---

the sacrifices and achievements of the ancestors that the clan *exists* at all … What can one give back to them? Sacrifices … does one ever give them enough?" (GM II: 19).

[117] As Reginster notes, "Emphasizing indebtedness toward God, as some commentators propose to do, will not help. If the feeling of indebtedness itself by no means decreases my worth as a person, it is hard to see how making it indebtedness toward God could have this effect" (2011: 67).

Nietzsche's account is incomplete because it does not offer a satisfactory solution to this problem.

I believe Nietzsche does provide a solution, but it is easy to miss because he offers it only in passing. Though initially the gods/ancestors were a projection of the tribe's fear and insecurity, they eventually became anthropomorphic projections of the virtues or qualities the tribe admired in themselves. He claims this occurred during the "middle period" of history, when "the noble clans [took] shape … who in fact returned, with interest, to their originators, the ancestors (heroes, gods) all of the qualities that had in the meantime become apparent in them, the *noble* qualities" (GM II: 19).[118] At the first stage, then, violating a custom was still only an occasion for *disapproval* from one's gods, grounded in a superstitious fear of punishment (see D 9), and reflected only the agent's indebtedness toward them. Being anthropomorphic projections of angry but powerful creditors, the failure to honor one's obligations to them, like the failure to repay one's debt to a human creditor, did not impugn the agent's sense of self-worth. At the second stage, however, when the gods exemplified the community's virtues and internalized standards of right conduct, the failure to honor one's obligations to them would have had greater significance. Indeed, I want to suggest this would have occasioned guilt, the "feeling of personal obligation" (GM II: 8).

To make sense of this, I hope the following analogy will prove helpful. Consider how a child disappointing a parent can elicit a different reaction than disappointing another authority figure (e.g., the school principal). Assuming that the parent serves as

[118] He expands on these remarks in the *Antichrist*, especially sections 16-18 and 24-6. In Judaism he believes this noble kind of virtue projection predated the Second Temple Period (538 BCE to 70 CE), and that the "holy God" originated within Judaism in response to the Babylonian Exile. This also coincided with the "slave revolt" in morality (GM I: 7): "The same instinct which prompts the subjugated to reduce their god to the 'good-in-itself' also prompts them to eliminate all the good qualities from the god of their conquerors; they take revenge on their masters by turning their god into the *devil*" (A 17). For analysis, see Snelson (2017).

an embodiment of values and ideals the child is sincerely trying to emulate, it is plausible to suppose that, when she falls short of the parent's expectations, she would experience guilt and not simply fear of punishment. If, for example, a child lies to her mother, she must not only deal with the consequences of lying (her punishment), but with the remorse she would feel for having disappointed her mom, for failing to live up to the standard of honesty that her mother instills and exemplifies. On the other hand, lying to a principal would only invoke anxiety and fear about the possible punishment. Failing to repay a debt (honor an obligation) to a god that is a projection of virtues and standards of right conduct that one accepts, I am suggesting, is like lying to one's parents, whereas failing to repay a debt to a god who is merely a projection of fear and anxiety is like lying to one's principal.

I am not trying to claim that what we have uncovered is a terribly mature notion of guilt at this point. After all, we might question whether the child would feel guilty if her mother would not be disappointed, which is just to say we can doubt whether she really holds herself to moral expectations that would impugn her sense of worth, or whether these are just the expectations of an admired authority she wants to please.[119] The same can be said for the moral agent now under consideration. Still, any developmental analysis of a natural phenomenon will involve these kinds of gray areas. It will require that the phenomenon emerge on a continuum and in degrees, over long periods of time, clearest perhaps only at the margins, and possibly even then only dimly.[120] Like a developing child's moral compass, guilt did not emerge on Nietzsche's analysis fully developed from the womb of bad conscience$_m$. With this caveat in mind, I

[119] The violation of which would result in withdrawal of parental affection, as Freud emphasizes.

[120] For instance, the phenomenology of guilt remains "dim." It is often difficult to parse the feeling of guilt from emotions like fear and shame.

hope it is plausible to suppose that what this agent experiences is a lot closer to guilt than fear of punishment.

If this analogy sticks, we can make sense of the appearance of *Schuldgefühl* in GM II: 19. Guilt feelings are *not* the consequence of our becoming indebted to more powerful gods, as Nietzsche's confusing remarks suggest. Rather, *Schuldgefühl* are the consequence of becoming indebted to gods who represent and exemplify standards of personal worth we accept and try to emulate. So, when we fail to honor our obligations to them, we experience guilt, "the feeling of personal obligation" (GM II: 8). And now our moral agent is starting to come into view: this agent *hold himself to expectations*, because he undergoes episodes of bad conscience$_m$, and those expectations are *moral*, as opposed to merely prudential or conventional, because when he violates them he judges his doing so to be "bad," and his self-punishment to be merited or deserved, in reference to a conception he has of *himself* as being either "good" or "bad." What he experiences is therefore the "meta-affect" of guilt, not just the "basic affect" of raw guilt.

Note, guilt does require judging one's *action* to be "bad," and therefore for bad conscience$_m$ to be merited or deserved in response to violating personal obligations. At the very least, it involves the implicit judgment that internal suffering is merited or deserved as repayment for one's action, for *what one did*. Connecting back up to the notion of debt, guilt is in this way *the internalization of the state of indebtedness*. Guilt is the emotion whereby the agent's awareness of the fact that he deserves punishment in the eyes of *others* for having reneged on an obligation, his being in a state of objective guilt, becomes a judgment that he makes against *himself*, thereby becoming the state of subjective guilt. I believe it is a further question, however, whether one ought to also judge *oneself* "bad" or "evil" because one has violated a personal obligation. Guilt always

"engages" our sense of worth as a person, as Reginster (2011: 66) put it, because it is the feeling of person obligation. But it can engage our self-worth in two ways. Guilt can manifest as guilt for *what I have done*—my action—or as guilt for *who I am*—what my action reveals about who I am as a person, or my character. These two judgments frequently overlap, but they are in fact separable. For instance, I might negligently make some remark that hurts your feelings, and seeing this I feel bad and apologize. But my undergoing an episode of bad conscience$_m$ in this way need not involve the further judgment that I am a bad *person* because I happen to be negligent at times. This would only follow if I believed, say, to be a good person, I must be *infallible* (or close to it).

I will now expand on these observations in connection to Nietzsche's analysis of ancient Greek guilt. As we will see, they used their gods to manage bad conscience$_m$ in an effective and healthy manner, in an effort to resist the judgment that they were "bad," even though they judged their actions to be so. In other words, they felt and accepted guilt for *what they did* but they refused to accept guilt for *who they were*.

4.2 Greek Blame and Responsibility

Nietzsche contrasts Greek religion with Christian religion after explaining how *Schuld* becomes "moralized" through Christianity. The obvious implication is that, among the ancient Greeks, *Schuld* did not become "moralized." This passage is therefore important to understanding what "moralization" consists in. As I argued above, it is not the process whereby material indebtedness becomes subjective guilt. I will provide further evidence of that argument here, by showing that the ancient Greeks did suffer from episodes of bad conscience$_m$ that they interpreted as guilt. However, I believe this has been obscured by the fact that this passage can also plausibly be interpreted as suggesting the

Greeks did not feel guilt at all, but only shame. One of my aims here is to show that this stronger reading is mistaken.

Nietzsche's main contention in this passage is that the Greeks used their gods to "keep 'bad conscience' at arm's length," by attributing their immoral actions to the gods, thereby making them take on the agent's "guilt" (GM II: 23). This can be interpreted in a stronger or weaker sense:

(i) The Greeks did not feel guilt, because they projected bad conscience$_m$ onto their gods, or
(ii) The Greeks did feel guilt, because they suffered from episodes of bad conscience$_m$, but they nonetheless (for some reason) projected their guilt onto the gods.

The stronger reading (i) implies that the Greeks did not experience guilt because they projected their "raw" guilt, their desire to punish themselves, along with any feelings of remorse, onto the gods. This is not an implausible view, because it is commonly believed and claimed that the Greeks lived in a shame culture, not a guilt culture.[121] But is this consistent with what Nietzsche says in this passage?

> "A god must have beguiled him," he said to himself finally, shaking his head … This way out is *typical* of the Greeks … In this manner the gods served in those days to justify humans to a certain degree even in bad things, they served as causes of evil—in those days it was not the punishment they took upon themselves but rather, as is *more noble*, the guilt … (GM II: 23)

I take it this means that the Greeks used their gods to either *excuse* or *exempt* themselves ("to a certain degree") from blame and responsibility. Again, this can be interpreted in a stronger and a weaker sense:

(iii) The Greeks did not feel responsible, because they blamed the gods for their evil actions, and so did not experience guilt, or
(iv) The Greeks did feel responsible, despite blaming the gods for their evil actions, and so did experience guilt.

[121] See Williams (1993). Williams himself thinks there is something to this idea, though he argues that the Greeks did accept responsibility for their actions. On my reading of Nietzsche, this would mean they felt guilt.

Note, (i) and (iii) are complimentary in the following way: The *reason* the Greeks did not feel guilty or responsible (iii) is that they did not experience bad conscience$_m$ (i). They never felt "pangs of [bad] conscience," never felt guilt, remorse, or contrition for what they did. This reading is at least clear, whereas the weaker reading is mysterious, if not incoherent. Why would the Greeks feel guilty and responsible for their evil actions, if they believed the *gods were to blame* for those actions? Nevertheless, I will argue that this weaker reading is correct, and that the stronger reading is inconsistent with what Nietzsche concludes in the passage.

First, note that if the Greeks never experienced episodes of bad conscience$_m$, there would be no desire or urge to blame the gods for their misdeeds. Ok, but the proponent of the stronger reading might reply: The wrongdoer is not projecting his guilt onto the gods because he feels *blameworthy*; he projects it because *others* are blaming him and he is trying to "excuse" himself from their resentment, to show that it is unjustified. Fair enough. We must then make sense of the fact that the Greek gods took on the agent's "guilt," "as is *more noble*," but "not the punishment." Because the gods did not take on the punishment, that means the *agent* took on the punishment. But again, why would he accept punishment for what he did if he did not feel *responsible* for his action? This is indeed very puzzling, and the stronger reading cannot explain the agent's willingness to do so. I think Bernard Williams' discussion of Agamemnon in *Shame and Necessity* provides some clarity.

First, some pertinent background: Agamemnon has just stolen Briseis from Achilles, but now must apologize because he needs Achilles to defeat the Trojans. According to Williams, "What [Agamemnon] suggests is that when he had that intention [to steal Briseis], he was in an abnormal state of mind, and this state of mind had a

supernatural explanation," yet, this "does not mean that it is not [Agamemnon's] business to make up for it" (1993: 53). Agamemnon reasons as follows: "But since I was deluded and Zeus took my wits away from me, I am willing to make all good and give back gifts in abundance" (quoted in Williams 1993: 53). Again, this is very confusing, but Williams explains: "[Agamemnon] is not dissociating himself from his action; he is, so to speak, dissociating the action from himself" (1993: 54). In other words, he accepts responsibility *for what he did*, but he does not take responsibility for *who he was* (at the time). In the former sense, "he does accept responsibility" (Williams1993: 53); in the latter sense, he does not. And this is also what Nietzsche concludes about the Greeks: "'foolishness,' 'lack of understanding,' a little 'disturbance in the head,' this much even the Greeks of the strongest, bravest age *allowed* themselves as the reason for much that was bad and doom-laden:—foolishness, *not* sin! Do you understand that?" (GM II: 23)

What Nietzsche is suggesting, I take it, is that the Greeks did accept responsibility for their *actions*, though they refused to accept guilt or responsibility for *who they were*, or for their character, when they performed heinous and evil actions, because they saw themselves as "men of noble descent, of happiness, of optimal form, of the best of society, of nobility, of virtue" (GM II: 23). In other words, they refused to indulge in the notion that they were guilty for being fallible, human, and overcome by passion, because they had such a high opinion of themselves, and so the idea that they could be *sinful*, as opposed to simply being "foolish" or "disturbed," could not have been more foreign to them. So, when Agamemnon stole Briseis, he recognized that action as immoral and he recognized that he did it, as something he was responsible for, but because he did it in a moment of passion (once Zeus had "stolen his wits"), he did not take it to be indicative of his character, of his "real" or "true" self. And the underlying reason for this is that he

had a *good opinion* of himself; he still judged himself to be "good." He judged himself to be good even though he recognized that *what he did* was "bad," and though he thought compensation or punishment was merited or deserved for what he did.

We can analyze this case somewhat awkwardly in terms of Strawson's framework. Interestingly, Agamemnon's rationale presumes that he did not believe he was *excused* from blame, because the intention he acted upon did display ill will. He just didn't think that intention was *his*, but rather Zeus' intention implanted in him (or his intention while overcome by rage). Also, Agamemnon did not exempt himself from blame for *what he did*, because he accepted his punishment. Rather, he took himself to be exempt from blame for *who he was* (at the time). That is, he thought he ought to apologize and make amends, which presumes that the obligation he violated was a personal or moral obligation. But he should *not* have to apologize for who he is as a person, given that he was not himself in that moment.[122] So, the gods did serve to justify the ancient Greeks "to a certain degree" (GM II: 23), namely, the gods exempted them from feeling the enduring kind of guilt that attaches to us as *persons*, or attaches to our character.

In light of the above observations, let's return to Clark's analysis of retributive punishment. She argued that retributive ideas work to isolate the criminal from his deed, so that he may remain a member of the community by repaying his debt through punishment. This was important to her Strawsonian reading of GM II, as it implied that we were no longer taking the "objective attitude" toward offenders, unlike exile punishment. That is what is going on here, with one important difference. Punishment

[122] Relatedly, Strawson mentions as considerations of *temporary* exemption things like "He wasn't himself" and "He has been under very great strain recently" (1967: 78). Strawson does not explicitly entertain the possibility that we might still take (partial) responsibility for our actions in these cases, but nothing I can see rules it out, and either or both of these considerations could plausibly apply to Agamemnon in this case. (He was livid over Achilles' previous insubordinate acts and in the midst of a war with the Trojans.)

is not meted out by the community—or, in this case, Achilles—but Agamemnon himself. He accepts the consequences of his actions, judging those to be "bad," and makes the necessary amends, and in that sense *takes responsibility* for what he did. If this is right, he is unlike the psychopath whom we considered at the end of Section 2. For Agamemnon, the prohibition against stealing another man's “property” is not a custom or conventional rule. It is a personal obligation that bears on his sense of worth, because his idea of the good person contains within it the idea "one who does not steal." He therefore holds himself to *moral* expectation not to steal; he is a moral or responsible agent.

A final consideration speaks in favor of the weaker reading. Nietzsche does not advocate deflecting responsibility, but just the opposite. "Signs of nobility: never thinking of degrading our duties into duties for everybody; not wanting to delegate, to share, one’s own responsibility; counting one’s privileges and their exercise among one’s *duties*" (BGE 272). That is, he advocates and admires *taking* responsibility. So, presumably Nietzsche would not advocate deflecting blame—when blame is warranted—and so long as it is blame for what one has done. That is the lesson we learn through the ancient Greeks. They came up with a rather ingenuous solution to ensure that their guilt stopped at guilt for *what they did*, by projecting guilt for *who they were* onto their gods.

I take it the following aphorism supports this lesson, and more importantly provides confirmation of the fact that the ancient Greeks did experience guilt on Nietzsche’s analysis. The problem with Christian or “moralized” guilt, as the passage makes clear, is that it goes far beyond being guilt for what one has done.

> *Justice which punishes*. – "Misfortune and guilt" – Christianity has placed these two things on a balance: so that, when misfortune consequent on guilt is great, even now the greatness of guilt itself is still involuntarily measured by it. But this is not *antique*, and that is why Greek tragedy, which speaks so much yet in so

> different a sense of misfortune and guilt, is a great liberator of the spirit in a way in which the ancients themselves could not feel it. They were still so innocent as not to have established an "adequate relationship" between guilt and misfortune. The guilt of their tragic heroes is, indeed, the little stone over which they stumble and perhaps break an arm or put out an eye: antique sensibility commented: "Yes, he should have gone his way a little more cautiously and with less haughtiness!" But it was reserved for Christianity to say: "Here is a great misfortune and behind it there *must* lie hidden a great, *equally great* guilt, even though it may not be clearly visible! ... In antiquity there still existed actual misfortune, pure innocent misfortune; only in Christendom did everything become punishment, well-deserved punishment ... so that with every misfortune [the sufferer] feels himself morally reprehensible and cast out. Poor mankind!—The Greeks have word for indignation at another's unhappiness: this affect was inadmissible among Christian peoples and failed to develop, so that they also lack a name for this *more manly* brother of pity. (D 78)

V. Perverse Guilt ("Moralized" Guilt)

As we saw above, Christian or "moralized" guilt manifests as a "will to find *oneself* guilty" (GM II: 22, emphasis mine), and is created to explain our "absolute unworthiness" as a consequence of uninhibited bad conscience$_m$. This kind of guilt comes into the world through the same general process as non-moralized guilt: by a community projecting its virtues and internalized standards of right conduct onto its gods, which come to represent *ideals* of those virtues and standards that individuals then measure themselves against. The crucial difference is that, once guilt becomes "moralized," it becomes an enduring state of the person that cannot be paid off.

As Nietzsche says, moralization occurs when the concepts of guilt and duty are "pushed back into conscience, more precisely, the entanglement of *bad* conscience with the concept of god" (GM II: 21). This formulation is not terribly clear, but what he means is that moralization occurs once the concepts of guilt and duty become attached to an idea of God who is the product of bad conscience$_m$, and hence a God who represents standards of personal worth or virtue that can never be realized by human beings. This explains why "moralized" guilt is an *inexpiable and enduring state of persons*, and does

not stop at being guilt for one's actions. As I noted previously, “moralized” guilt perverts an already extant notion of guilt. Here I will defend that claim by showing that “moralized” Christian guilt is a perversion of Jewish guilt.

5.1 Jewish Guilt

Nietzsche believed that, prior to the Babylonian Exile, Jewish monotheism mirrored Greek polytheism in many respects. In particular, the Jews’ concept of Yahweh, like the Greek gods, was a projection of the Israelites' "good conscience":

> Originally, especially at the time of the kings, Israel stood in the right, that is, the natural relationship to all things. Its Yahweh was the expression of a consciousness of power, of joy in oneself, of hope for oneself: through him victory and welfare were expected; though him nature was trusted to give what the people needed—above all, rain. Yahweh is the god of Israel and therefore the god of justice: the logic of every people that is in power and has a good conscience. (A 25, see also A 16)

The problem is that, as anyone who is familiar with the history of Judaism knows, this period of prosperity did not last. The Jews first dealt with internal turmoil and the fracturing of their nation state into the kingdoms of Judah and Samaria in 931 BCE. Samaria was then conquered by the Assyrians in 722 BCE. The kingdom of Judah survived, but was conquered by the Babylonians in 587 BCE. From that point until the Common Era, the Jews lived exiled from the land Yahweh had promised them, under the Babylonians, the Greeks, and the Romans, respectively.[123]

Relying on his training as a philologist, Nietzsche dissected the biblical texts written during this Second Temple period (538 BCE to 70 CE), and he (and other biblical scholars) concluded that the priestly elements within Judaism rose to power during this

[123] See Snelson (2017) for analysis.

time.[124] The priests convincingly interpreted this series of misfortunes as evidence of their guilt, as a state of punishment inflicted by Yahweh:

> But all hopes remained unfulfilled. The old god was no longer able to do what he once could do. They should have let him go. What happened? They change his concept … at this price they held onto him. Yahweh the god of "justice"—no longer one with Israel, an expression of the self-confidence of the people … The concept of God becomes a tool in the hands of priestly agitators, who now interpret all happiness as a reward, all unhappiness as punishment for disobeying God, as "sin." (A 25)

> These priests accomplished a miracle of falsification, and a good part of the Bible now lies before us as documentary proof … they translated the past of their own people into religious terms, that is, they turned it into a stupid salvation mechanism of guilt before Yahweh, and punishment; of piety before Yahweh, and reward. (A 26)

The priests created a "moral world order," according to which "the value of a people … is to be measured according to how much or how little the will of God is obeyed" (A 26).

Clearly this later conception of Yahweh is a projection of bad conscience$_m$. To be precise, the Jews interpreted their exile as a consequence of their indebtedness, as their being judged guilty by god, which served as an explanation of their misfortune and subjugation (the latter being the actual cause of their bad conscience$_m$). By interpreting their misfortune in this way, their bad conscience$_m$ was interpreted as a consequence of the failure to live up to holiness standards of personal worth or virtue the priestly elements within Judaism widely accepted. The Jews who then accepted this explanation thereby also accepted the divine authority of the holiness covenant they espoused. The Jews then came to see the priestly holiness norms as personal or moral obligations.

Exilic Jewish guilt differs from Greek guilt in three crucial respects. First, the Greek gods did not sit in judgment of their people, or condemn them for their immoral actions. As Nietzsche says in *Human, All too Human*:

---

[124] Notably Julius Wellhausen, arguably the most influential Old Testament scholar of the 19th Century.

> The Greeks did not see the Homeric gods as set above them as masters, or themselves set beneath the gods as servants, as the Jews did. They saw as it were only the reflection of the most successful exemplars of their own caste, that is to say an ideal, not an antithesis of their own nature. They felt inter-related with them, there existed a mutual interest, a kind of symmetry. Man thinks of himself as noble when he bestows upon himself such gods … (HA I: 114)

The reason for this is that, secondly, Jewish guilt manifests as indebtedness toward a God that is a projection of bad conscience$_m$, whereas the Greek gods were not.[125] Finally, this greater degree of strength is evident by the fact that the Jews blamed themselves for *who they were* and not only for *what they had done*. Specifically, this occurred on Nietzsche's analysis when Jewish guilt became attached to the concept of sin, and the Jewish moral law became a holiness covenant, as opposed to a warrior ethos.

Importantly, however, on Nietzsche's analysis Jewish guilt remained in principle expiable. The Jews believed they could *repair* their relationship with God by honoring the new holiness covenant. Although Yahweh *now* sat in judgment of them, they did not believe that he would *forever* sit in judgment of them. Nietzsche takes this to be evident from the fact that the Jews still believed themselves to be God's "chosen" people, to be the unique and singular object of his affection, because they were capable of being "holy" just like him—*if* they fulfilled the new law. Accordingly, their idea of the "holy God" retained the defining feature of their nobility—their holiness—it remained a projection of virtues they were uniquely capable of exemplifying.[126] As evidence of this, consider the quote below where Nietzsche contrasts Christianity with Judaism's concept of a good person:

---

[125] The explanation for this is that, as noted in Section 3, bad conscience$_m$ is realized in *degrees*, according to the strength and/or duration of a group's subjugation. So we are to infer that bad conscience$_m$ became more pronounced in the Jews than in the Greeks.

[126] More accurately, their conception of "God" is still noble because it is an expression of their "pathos of distance" (GM I: 2), their feeling of superiority, which is the basis of noble valuation. See Snelson (2017).

> The "holy people," who had retained only priestly values, only priestly words for all things and who, with awe-inspiring constancy, had distinguished all powers on earth from themselves, as "unholy," as "world," as "sin"—this people produced an ultimate formula ... that was logical to the point of self-negation: as *Christianity*, it negated even the last form of reality, the "holy people," the "chosen people." (A 27)

As we will now see, Christianity "negates" *any* sense of nobility, and part of the reason for this is that it understands guilt to be a condition of not only *who one is*, but who one is and *must* be. A consequence of this is that it also makes a virtue out of the feeling of guilt, as something it is "good" to experience because one is necessarily "bad." This, of course, is in stark contrast to Greek guilt, but also Jewish guilt.

5.2 Christian Guilt

According to Nietzsche, Christianity co-opted Judaism's "holy God" and its concept of "sin" as an explanation for personal suffering. However, Christian guilt goes far beyond Jewish guilt, because it represents "the *will* of man to find himself guilty and reprehensible to the point that it cannot be atoned for; his *will* to find himself punished without the possibility of the punishment ever becoming equivalent to the guilt" (GM II: 22). As Nietzsche says in *Human, All too Human*, "Christianity ... crushed and shattered man completely and buried him as though in mud: into a feeling of total depravity (HA I: 114). The reason for this is not simply that Christians suffered from *ressentiment* and bad conscience$_m$ on Nietzsche's analysis, but because their concept of the "holy God," unmoored from the Jewish holy law that served as a gateway to divine states of being, only serves to represent an ideal of holy perfection humans can never achieve. Christianity's moral law, as Nietzsche characterizes it in *Daybreak*, is "the canon of *impossible virtue*":

> In the New Testament, the canon of virtue, of the fulfilled law, is set up: but in such a way that it is the canon of *impossible virtue:* those still *striving* after morality are in the face of such a canon to learn to feel themselves ever *more*

> *distant* from their goal, they are to *despair* of virtue, and in the end *throw themselves on the bosom* of the merciful – only if it ended this way could the Christian's moral effort be regarded as possessing any value, with the presupposition therefore that it always remains an unsuccessful, miserable, melancholy *effort*. (D 87)

The Jewish moral law, by contrast, is only a canon of *unfulfilled* virtue. Judaism can in principle dispense with guilt for *who one is*, whereas Christianity cannot.

Jewish guilt and Christian guilt are both more severe than Greek guilt because they take as their primary object the person, or the person's character, and not just her actions. Also, I think we have good reason to think Nietzsche would have rejected this additional judgment of guilt, because in both cases it rests on the spurious idea of sin. That said, Christian guilt is clearly perverse in a way Jewish guilt is not. It is perverse because it assumes we are guilty for simply being born within, and being part of, the natural world we inhabit. Christian guilt attaches to *all* humans, throughout time, who *always* fall short of the standard of personal worth exemplified by its "holy God":

> [T]hink here of the *causa prima* of man, of the beginning of the human race, of its progenitor, who is now burdened with a curse ("Adam," "Original Sin," "unfreedom of the will") of nature, from whose whom womb arises and into which the evil principle is now placed ("demonizing of nature") or of existence generally, which is left *valueless in itself*. (GM II: 21)

At the end of this quote Nietzsche is alluding to the fact that the Christian god is an expression of the "ascetic ideal," because worshipping this god involves, implicitly or explicitly, judging our lives in this world to be disvaluable, unless we "negate" this world by valuing it as a means to the good life to come.[127] Again, though the Jewish conception

[127] The ascetic "relates our life (together with that which it belongs: 'nature,' 'world,' the entire sphere of becoming and of transitoriness) to an entirely different kind of existence, which it opposes and excludes, *unless*, perhaps, it were to turn against itself, *to negate itself*: in this case, the case of an ascetic life, life is held to be a bridge for that other existence. The ascetic treats life as a wrong path that one must finally retrace back to the point where it beings; or as an error that one refutes through deeds—*should* refute: for he *demands* that one go along with him; where he can, he forces *his* valuation of existence" (GM III: 11).

of god came to be ascetic in many respects, worshipping it never involved making this ascetic *evaluation* of life.[128]

Consequently, unlike Greek guilt, "moralized" guilt is not simply guilt over *what one has done*; unlike Jewish guilt, it is not simply guilt for *who one is*; it is guilt for who one is *and must be*. In Christianity, guilt becomes a pervasive and inescapable feature of the human condition. Therefore, "moralization" is not Nietzsche's explanation of the process whereby indebtedness becomes guilt, and in fact presumes that one already has a sense of "personal obligation." "Moralization" is rather the process whereby *Schuld* and *Pflicht* become attached to an ideal of personal worth that humans cannot possibly live up to, because these correspond to a conception of God that is a projection of bad $\text{conscience}_{\text{m}}$. From this point of view, we are categorically and forever "bad," "evil," or "sinful," and so the guilt we experience is always *good*. As Christopher Janaway says:

> Moralization is the elevation of feeling guilty into a virtue, its incorporation into what the morally good individual is or does, into a conception of the kind of person one should want to be, by means of the rationalizing metaphysical picture in which the individual's essential instinctual nature *deserves* maltreatment, because it stands in antithesis to an infinite creditor. (2007: 142)

VI. Is Guilt Something That Ought to be Overcome?

In closing, I'd like to consider a challenge to the interpretation of guilt I've offered here. Nietzsche frequently remarks that guilt is irrational and unhealthy, something that ought to be overcome. Below are two representative passages.

> *Towards the re-education of the human race.* – Men of application and goodwill assist in this one work: to take the concept of punishment which has overrun the whole world and root it out! There exists no more noxious weed! Not only has it been implanted into the consequences of our actions –and how repugnant to reason even this is, to conceive cause and effect as cause and punishment! – but

[128] The Jews remained to the end on Nietzsche's analysis a people "firmly attached to life—like the Greeks and more than the Greeks … these strange people … did not desire to get rid of their bodies but … hoped to retain them for all eternity" (D 72). Even the analysis he provides of the Jews in the *Antichrist*, which is critical of the efforts they took to remain attached to this world and find meaning within it, assumes that they never judged it to be *unworthy* of living.

> they have gone further and, through this infamous mode of interpretation with the aid of the concept of punishment, robbed of its innocence the whole purely chance character of events. Indeed, they have gone so far in their madness as to demand that we feel our very existence to be a punishment – it is as though the education of the human race had hitherto been directed by the fantasies of jailers and hangmen! (D 13)

> Men were considered "free" so that they might be judged and punished—so that they might become *guilty:* consequently, every act had to be considered as willed, and the origin of every act had to be considered as lying within consciousness … Today, as we have entered into the reverse movement and we immoralists are trying with all our strength to take the concept of guilt and the concept of punishment out of the world again, and to cleanse psychology, history, nature, and social institutions and sanctions of them, there is in our eyes no more radical opposition than that of the theologians, who continue with the concept of a "moral world order" to infect the innocence of becoming by means of "punishment" and "guilt." Christianity is a metaphysics of the hangman. (TI VI: 7)

Brian Leiter (2011, 2019b) has argued, on the basis of passages like these, that Nietzsche wants us to dispense with the feeling of guilt. I think the relevant question to ask is: *which* idea(s) of guilt does Nietzsche want us to overcome?

Leiter believes that we should overcome *all* notions of guilt because they rest on the mistaken belief in free will. As he says, "Once we abandon this 'error of free will' we should, in turn, abandon the reactive concepts whose ineligibility depends on it, concepts like 'guilt'" (Leiter 2011: 105). But this does not follow. For one thing, as I argued in Chapter 2, Nietzsche nowhere identifies free will as a condition of subjective guilt in GM II. Secondly, if my interpretation of Greek guilt is correct, they had a *compatibilist* notion of responsibility, because blame and responsibility were not taken to be unjustified despite the common belief that the gods determined their most heinous actions. So again, *which* idea(s) of guilt does he think we ought to overcome? Well, which idea(s) is he talking about above?

There can be no question that Nietzsche is objecting to Christian guilt and its "hangman metaphysics" in the second passage from *Twilight*, which he also thinks

requires belief in libertarian free will (GM I: 13). In the passage from *Daybreak* he is also objecting to Christian guilt, of individuals who "demand that we feel our very existence to be a punishment." However, both passages also implicate Jewish guilt, because it is "repugnant to reason" to "conceive cause and effect as cause and punishment." Recall, that is what the Jewish priests did during the Babylonian Exile, by interpreting their suffering as a deserved consequence of their sins. They "robbed misfortune of its innocence" (A 26). We can safely conclude, then, that Nietzsche is opposed to any notion of guilt that rests on the idea of sin. Because the idea of sin grounds the judgment that one is guilty for *who one is*, we can infer that he is at least skeptical of that form of guilt as well—*if* it necessarily depends on the idea of sin.

Yet we are still left with a problem. Greek guilt is healthy and worth emulating, I have suggested, because the Greeks invented an ingenuous solution to the feedback loop of self-inflicted violence that bad conscience$_m$ creates. If left unchecked, bad conscience$_m$ becomes guilt for *who one is*, and then guilt *who one must be*. The Greeks, as we have seen, resisted these latter judgments of guilt, by blaming their gods for their evil actions. However, I take it to be obvious that this expedient for mollifying bad conscience$_m$ is not viable for us today. So, to reframe our question: is it possible for guilt to be healthy for us today? Is it possible for guilt to stop at guilt for one's actions? And should it? Or might it manifest as guilt for who one is, in a way that does not presume the idea of sin?

Note, not all guilt is serious or weighty enough to leave a lasting mark, and in these cases it is certainly possible for guilt to stop at what one does. The guilt some claim to experience for reneging on their diets is plausibly like this (unless, perhaps, one has an eating disorder or is zealously dedicated to her health). Or, if I lose my temper and snap back at my wife out of anger, I will feel guilty afterward and apologize—will experience

an episode of bad conscience$_m$—but if this a seldom occurrence it will not register as guilt over who I am. My guilt would be like Agamemnon's guilt. However, if my snapping back is a frequent occurrence, and something I dislike about myself, then presumably it ought to register not only as guilt over *what I did* but as guilt over *who I am*. In this case what I did does not merely engage and temporarily damage my conception of myself as a good person; it leaves a lasting mark by pointing to a feature of my *character*. It seems I must judge myself to *be* a bad person, because my actions consistently demonstrate that I *am* someone I disapprove of and don't want to be. If I am to be a person of conscience, a person of integrity, as Nietzsche would encourage me to be (see BGE 272; D P 4; GM II: 2), Agamemnon's strategy is a non-starter. I cannot blame the gods for my anger, cannot blame my circumstances, or my emotions. I have to take responsibility for *myself*, and that would seem to involve accepting guilt for my character as constituted.

This reveals an important aspect of guilt that Leiter's recommendation neglects. Guilt is ugly, yes, because it involves maltreating oneself. But if I never beat myself up, that would mean I never held myself to moral expectations I sometimes transgress. And that means I would not be capable of becoming anything better than what I currently am, let alone something great, as Nietzsche would seem to want.[129] As Zarathustra advises, "And if a friend does you evil, then say: 'I forgive you what you did to me; but that you have done that to yourself—how could I forgive that?" (Z II: 3). Zarathustra recommends that we should forgive our friends of their debts, or no longer hold the guilt of their action against them. But by "forgiving" our friend we cannot and *should* not relieve him of the guilt that comes from violating his own standards, because that is something only he can do by taking responsibility *for himself*. Because this is something

[129] In fairness to Leiter, he only "neglects" this possibility because he doesn't think we can play any active role in constituting the self, as we discussed in Chapter 1.

Nietzsche consistently and emphatically implores us to do (see Chapter 1), I find it difficult to believe that he would want us to overcome guilt for who one is. That is, so long as one can do so without believing oneself to be sinful.

In closing, I would be remiss if I did not mention that Nietzsche often gives the impression that agents who experience guilt and remorse are necessarily "lesser" or "slavish." It is also true that the "great" and "healthy" human beings he often praises are not in the habit of apologizing. However, as Nietzsche himself was aware, accompanying this tendency is almost always an unrealistic and *inhuman* view of these people. For instance, when he praises Napoleon as one in whom "the ancient ideal itself stepped *bodily* and with unheard of splendor before the eyes and conscience of humanity," he is careful to note that Napoleon was the "synthesis of an *inhuman* and a *superhuman*" (GM I: 16). We should ask: Is he here praising Napoleon *for* his inhumanity, or *despite* his inhumanity? Or take the "sovereign individual" (GM II: 2), our focus in the next chapter. He is often portrayed as someone who would not feel guilt, because he is "above" morality. There is something to this, as I will argue. However, he often gets portrayed in this way because scholars stress that he is so well-ordered that he never makes promises he can't keep, because, it turns out, he is *infallible* (or nearly so).

This is *not* how the sovereign individual should be interpreted. Somewhere between this seemingly infallible ideal of humanity, and the hopelessly fallible ideal of humanity promulgated by Christianity, lies a fallible but healthy middle ground. I have here sought to find and articulate that middle ground, as a space that all humans can and do occupy. In the next chapter we will consider how Nietzsche aspires for us to become more than responsible agents capable of feeling guilt.

References

Works by Nietzsche

A: *The Antichrist*. 1888. In W. Kaufmann, trans. and ed., *The Portable Nietzsche*. New York: Viking Press, 1954

D: *Daybreak*. 1881. R.J. Hollingdale, trans., Maudemarie Clark and Brian Leiter, ed. New York: Cambridge University Press, 1997.

GM: *On the Genealogy of Morality*. 1887. Clark, Maudemarie, and Swenson, Alan J, trans. Indianapolis: Hackett Publishing Company, 1998.

HA: *Human, All Too Human*. 1878. R.J. Hollingdale, trans. Cambridge: Cambridge University Press, 1996.

GS: *The Gay Science*. 1882/1887. W. Kaufman, trans. New York: Vintage Books, 1974.

TI: *Twilight of the Idols*. 1888. In W. Kaufmann, trans. and ed., *The Portable Nietzsche*. New York: Viking Press, 1954

Z: *Thus Spoke Zarathustra*. 1883-5. In W. Kaufmann, trans. and ed., *The Portable Nietzsche*. New York: Viking Press, 1954.

Other Works

Boehm, Christopher. 2012. *Moral Origins*. New York: Basic Books.

Clark, Maudemarie. 2015a. "Nietzsche's Contribution to Ethics." In *Nietzsche on Ethics and Politics*, 62-74. Oxford: Oxford University Press.

___. 2015b. "Nietzsche on Free Will, Causality, and Responsibility." In *Nietzsche on Ethics and Politics*, 75-96. Oxford: Oxford University Press.

___. 2001. "On the Rejection of Morality: Bernard Williams's Debt to Nietzsche." In *Nietzsche on Ethics and Politics*, 41-61. Oxford: Oxford University Press.

___. 1998. "Introduction" In *On the Genealogy of Morality*. Clark, Maudemarie, and Swenson, Alan J, trans. Indianapolis: Hackett Publishing Company.

___. 1994. "Nietzsche's Immoralism and the Concept of Morality." In *Nietzsche on Ethics and Politics*, 23-40. Oxford: Oxford University Press.

Clark, Maudemarie and Dudrick, David. 2012. *The Soul of Nietzsche's* Beyond Good and Evil." New York: Cambridge University Press.

Darwin, 1871. *The Descent of Man*. London: Penguin Books.

De Waal, Frans. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge: Harvard University Press.

Freud, Sigmund. 1989. *Civilization and its Discontents*. In Peter Gay ed., *The Freud Reader*. New York: W.W. Norton and Company, pp. 722-772.

Greenspan, P.S. 1992. "Subjective Guilt and Responsibility." *Mind*, 101 (402): 287-303.

Janaway, Christopher. 2007. *Beyond Selflessness*. Oxford: Oxford University Press.

Korsgaard, Christine. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

Leiter, Brian. 2015. *Nietzsche on Morality*. London: Routledge.
___. 2013. "Moralities are a Sign-Language of the Affects." *Social Philosophy & Policy* 30, 237-58.
___. 2011. "Who is the 'Sovereign Individual'? Nietzsche on Freedom." In Simon May ed., *"Nietzsche's On the Genealogy of Morality": A Critical Guide*. Cambridge: Cambridge University Press.

Morrisson, Iain. 2018. "Nietzsche on Guilt: Dependency, Debt, and Imperfection." *European Journal of Philosophy*: 974-990.
___. Unpublished Manuscript. "Chronology in *On the Genealogy of Morality II*."

Reginster, Bernard. 2017. "What is the Structure of Genealogy of Morality II?" *Inquiry* 61 (1), 1-20.
___. 2011. "The Genealogy of Guilt." In May, Simon ed., *Nietzsche's "On the Genealogy of Morality": A Critical Guide*. Cambridge: Cambridge University Press, 56-77.

Ridley, Aaron. 1998. *Nietzsche's Conscience*. Ithaca: Cornell University Press.

Risse, Mathias. 2001. "The Second Treatise in *On the Genealogy of Morality*: Nietzsche on the Origin of Bad Conscience." *European Journal of Philosophy* 9 (1): 55-81.

Russell, Paul. 2011. "Moral Sense and the Foundations of Responsibility." In R. Kane ed., *The Oxford Handbook of Free Will: Second Edition*. Oxford: Oxford University Press, 199-220.
___. 2004. "Responsibility and the Condition of Moral Sense." *Philosophical Topics* 32: 287-305.

Snelson, Avery. 2017. "The History, Origin, and Meaning of Nietzsche's Slave Revolt in Morality." *Inquiry* 60: 1-30.

Strawson, P.F. 1962. "Freedom and Resentment." In G. Watson ed., *Free Will: Second Edition*. Oxford: Oxford University Press, pp. 72-93.

Taylor, Gabrielle. 1985. *Pride, Shame, and Guilt*. Oxford: Oxford University Press.

Tomasello, Michael. 2016. *A Natural History of Morality*. Cambridge: Harvard University Press.

Vincent, Ring, Rebecca, Sarah, and Andrews, Kristin. "Normative Practices of Other Animals." In A. Zimmerman, K. Jones, and M. Timmons, eds., *The Routledge Handbook of Moral Epistemology*. New York: Routledge, pp. 57-83.

Velleman, David J. 2003. "The Genesis of Shame." *Philosophy and Public Affairs* 30.1: 27-52.
___.1999. "The Voice of Conscience." *Proceedings of the Aristotelian Society* 99: 57-76.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

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

Chapter 5

Permitted Promising, Trust, and Sovereign Trust

Abstract: This concluding chapter analyzes the "highest form" (GM II: 3) of conscience that Nietzsche attributes to the "sovereign individual." This enigmatic figure is profiled in the second aphorism of GM II, where he is identified as the animal "das *versprechen darf*," literally one that "may," is "permitted," or "allowed" to promise. I argue that having the permission to promise is a metaphor for trust, which is an attitude of reliance from the participant stance (Holton 1994), and so involves regarding the promiser as a responsible agent. Trust is underwritten by the form of conscience Nietzsche calls the "memory of the will" (GM II: 1), which is an ability to extend practical commitment in the absence of external incentives. However, Nietzsche's remarks in the beginning of GM II invite conflicting interpretations regarding the nature of trust and this form of conscience. In the first aphorism he gives the impression that anyone with a memory of the will—presumably all humans—is permitted to promise, but in the second aphorism he identifies this person in the singular as the sovereign individual, because conscience is his "dominant instinct" (GM II: 2). This creates a problem concerning the *scope* of Nietzsche's genealogy of responsible agency. Are *all* humans permitted to promise, and thus sovereign individuals? Or are sovereign individuals rare, and thus most humans are *not* permitted to promise? My aim in this chapter is to offer a solution to this problem, by arguing that Nietzsche has in mind two different notions of trust, of what it means to be "permitted to promise." The weaker notion I argue corresponds to *moral trust*, to be distinguished from mere reliance. Moral trust presumes that one who is relied upon is a morally responsible agent and will keep her promises out of a sense of moral obligation. Moral trust, I argue, is underwritten by our capacity to feel guilt. The stronger notion corresponds to a much more demanding *ideal* of responsibility, tied to Nietzschean autonomy, which the sovereign individual exemplifies. I argue that he keeps his promises because he is committed to his *integrity*, whereas most of us do not keep our promises from a motive of integrity. A consequence of this is that most humans are not *trustworthy* on Nietzsche's analysis, though this fact does not impugn moral trust.

Keywords: sovereign individual, memory of the will, permitted promising, trust, guilt, integrity.

I. The Sovereign Individual and the Scope Problem

Apart from determining the precise point when debt becomes guilt, there is no greater lacuna in GM II's genealogy of responsibility than the role and place of the sovereign individual therein. This enigmatic figure is foreshadowed in the second aphorism and identified as the "ripe fruit" who emerges from the morality of custom (GM II: 2, 3). Nietzsche heaps so much praise upon him that it borders on the ridiculous. He is

claimed to be "autonomous," the "lord of the *free* will," and has achieved "mastery over himself" (GM II: 2); he alone knows the "extraordinary privilege of *responsibility*" (GM II: 2); he is even declared to represent the "feeling of the completion of man himself" (GM II: 2). As we can see, he has *many* superlative qualities. His defining feature, however, is that he is the person "das *versprechen darf*," literally one that "may promise," is "permitted to promise," or "allowed to promise."[130] As we have seen, GM II is also proclaimed by Nietzsche to be "the long history of the origins of *responsibility*," and he equates this with the "task" of "breeding an animal that is permitted to promise" (GM II: 2). So, being a responsible agent and having the permission to make promises are understood by him to be related concepts, if not synonymous.

Defining responsibility in terms of having the permission to make promises is strange to be sure, but it is also an incredibly suggestive formulation of responsible agency. It is so primarily because it has the potential to shed light on aspects of responsibility typically neglected or obscured by focusing on *moral* responsibility, and which are every bit as important from a naturalistic point of view. When we analyze responsible agency in terms of moral responsibility, we take it to be something like a status or regard that applies singularly to human beings who possess certain capacities to act (i.e.,, free will), or alternatively, in terms of certain capacities for rational or moral self-governance (i.e., conscience). We are "responsible" for our actions, meaning that we are appropriate targets of blame and the reactive attitudes, *because* we possess these

[130] Walter Kaufmann translates "das *versphrechen darf*" as the "the right to make promises," and Douglas Smith translates it as "entitled to make promises." Clark and Swenson translate *dürfen* as "permission." I prefer their translation for the reasons they cite (1998: 139), because "permission" captures the normative character of *dürfen* without introducing more complex normative ideas corresponding to rights and entitlements, but also and more importantly because the metaphor of "permission" invites an ambiguity that proves felicitous to my aims here. Giving another the "permission to promise" is a natural characterization of what we do when we *accept* their promise on the basis of moral trust, and, as we will see, the sovereign individual is "permitted" to *make* promises because he is trustworthy.

capacities. We have taken the latter approach to understanding responsibility thus far, understanding responsible *agency* to depend on the formation of bad conscience and the ability to feel guilt. When we think of responsibility in these terms, bad conscience and guilt demarcate human beings, as responsible agents, from other creatures in the natural order of things. In Strawsonian terms, guilt and bad conscience divide those agents who are *exempt* from the moral reactive attitudes from those who are not.

The opening lines of GM II confirm that Nietzsche is interested in the issue of exemption, and that, moreover, possessing the status of a permitted promiser is tracking this very distinction. The essay begins with him asking, "To breed an animal that is *permitted to promise*—isn't this precisely the paradoxical task nature has set for itself with regard to man? Isn't this the true problem *of* man? …" (GM II: 1). As Christa Davis Acompora observes, "Clearly, this is a question that is raised about humankind generally. It applies to the kind of being that makes us human beings" (2006: 15). Bernard Reginster agrees:

> The whole of the Second Essay is framed by a fundamental distinction between the 'the prehistory of mankind,' in which the 'human *animal*' is subjected to forces that transform him into 'an animal *with the right to make promises,'* at which point we enter 'history' proper, which features 'mankind come to completion.' Acquiring the right to make promises distinguishes the human being from the rest of the animal realm, which includes his 'pre-historic' self. (2017: 7)

So, according to the opening lines of GM II, having the permission to make promises is a status that demarcates human beings *in general* from other kinds of animals. In Chapter 4, we saw that our achieving this status depends on the formation of bad conscience, which resulted in humanity's "forceful separation from [our] animal past" (GM II: 16). Bad conscience made this transformation possible by giving depth to the human soul,

enabling us to hold ourselves to moral expectations, and to eventually experience guilt, the "feeling of personal obligation" (GM II: 8), when we violated them.

Note, however, that having the permission to promise seems *not* to be concerned, at least not obviously, with whether humans possess a bad conscience and are able to experience guilt. Nor does it seem to be concerned with whether they possess free will, or even capacities for rational or moral self-governance such that it would be appropriate to praise and blame them. That is, Nietzsche seems not to be focused on *moral* responsibility, but on responsibility of another—though perhaps related—type. Perhaps it will turn out that having the permission to promise depends on being a morally responsible agent who is susceptible to the moral reactive attitudes, but at least on the face it Nietzsche's concern seems to be elsewhere. On what, exactly? Over the past decade there has been a great deal of convergence in the secondary literature, to the effect that having the permission to promise is a metaphor for *trust*. I agree. However, scholars have offered divergent and conflicting interpretations about the nature of Nietzschean trust, and they have offered these accounts for the most part precisely in an effort to *deflect* the implication that Nietzsche's analysis of trust could in any way imply that humans are *morally* responsible. I believe this is a mistake, at least as it concerns one form of trust, as I will argue here.

Essentially the disagreement among scholars centers around two related issues that will frame this paper. First, what capacities are conferred by the form of *conscience* underwriting permitted promising, which Nietzsche calls the "memory of the will" (GM II: 1), and secondly, given this understanding of conscience, who qualifies as a sovereign individual? The first question concerns the nature and execution of the memory of the will, its significance for human action more broadly, and its relationship to trust; the

second question concerns the *scope* of Nietzsche's genealogy of responsible agency. In short, are most humans permitted to promise, and thus sovereign individuals? Or are sovereign individuals rare and only they permitted to promise? As we saw in Chapter 2, Nietzsche offers divergent and conflicting remarks concerning the scope issue, which in turn frames how we ought to understand the sovereign individual and the nature of sovereign conscience specifically.

We saw evidence of this ambiguity just a moment ago. In the first aphorism Nietzsche suggests that all humans are permitted promisers, because this is described as a kind of status that distinguishes humans from other animals. As further evidence of this he even says the "task" of "breeding" such an animal has "been solved to a high degree" (GM II: 1). The reason, presumably, is that having the permission to promise depends on possessing a "memory of the will," and Nietzsche gives us every impression to think that most humans have one:

> Precisely this necessarily forgetful animal in whom forgetting represents a force, a form of *strong* health, has now bred in itself an opposite faculty, a memory, with whose help forgetfulness is disconnected for certain cases—namely for those cases where a promise is to be made: it is thus by no means simply a passive no-longer-being-able-to-get-rid-of the impression once it has been inscribed, not simply indigestion from a once-pledged word over which one cannot regain control, but rather an active no-longer-wanting-to-get-rid-of, a willing on and on of something one has once willed, a true *memory of the will:* so that a world of new strange things, circumstances, even acts of the will may be placed without reservation between the original "I want," "I will do," and the actual discharge of the will, its *act*, without this long chain of the will breaking. (GM II: 2)

As I argued in Chapter 2, when Nietzsche characterizes conscience as a "memory of the will" he takes it to be an ability *to sustain practical commitment in the absence of external incentives*. The metaphor of conscience as the "will's memory" is meant to denote that conscience *just is* that which allows an agent to sustain the motivation necessary to carry an act of will through, understanding that motivation to be in some

sense internal to the person and self-perpetuating. We will analyze the memory of the will in more detail in Sections 3 and 4. The important thing to take note of here is that, according to GM II: 1, most humans have this ability and are therefore permitted promisers.

Nietzsche appears to contradict himself in the very next aphorism, however. There he identifies the sovereign individual in the singular as the "human being who is permitted to promise," and it's clear he has this status not simply because he *possesses* a "memory of the will." He is:

> [T]he individual resembling only himself, free again from the morality of custom, autonomous and supermoral (for 'autonomous' and 'moral' are mutually exclusive), in short, the human being with his own independent long will, the human being who is *permitted to promise*—and in him a proud consciousness, twitching in all his muscles, of *what* has finally been achieved and become flesh in him, a true consciousness of power and freedom, a feeling of the completion of man himself. This being who has become free, who is really *permitted* to promise, this lord of the *free* will, this sovereign—how could he not know what superiority he thus has over all that is not permitted to promise and vouch for itself, how much trust, how much fear, how much reverence he awakens—he "*earns*" all three—and how this mastery over himself also necessarily brings with it mastery over circumstances, over nature and all lesser-willed and more unreliable creatures? (GM II: 2)

The concluding sentence suggests that the sovereign individual "*earns*" trust because, in addition to possessing a memory of the will, he has achieved "mastery over himself." Presumably, and because of this, he also possesses the "proud knowledge of the extraordinary privilege of *responsibility*" (GM II: 2). As Lanier Anderson reminds us, "this kind of responsibility is explicitly supposed to be a 'privilege' earned only by a few" (2013: 438). In short, according to GM II: 2, most humans are *not* permitted promisers.

Call this *scope problem*; more precisely, it is a problem concerning the scope of GM II's genealogy of responsible agency—who is "permitted to promise?" Are all human

beings, and wouldn't' that imply that we are all sovereign individuals? Or are sovereign individuals rare and only they permitted to promise? Framed in these terms, the scope problem creates an interpretive dilemma, which we might explain in terms of sovereign conscience. If we accept the first horn of this dilemma, sovereign conscience is something that all humans possess. In that case it turns out that what Nietzsche says about permitted promising in GM II: 2 is confused: sovereign individualism and sovereign conscience are not *ideals*, but generally realized. If we accept the second horn of this dilemma, sovereign conscience is something that few humans possess. In that case it turns out that what Nietzsche says about permitted promising in GM II: 1 turns out to be confused: permitted promising is not a status that demarcates human beings from other kinds of creatures. Affirming either horn of this dilemma has the consequence of rendering Nietzsche's remarks in the opening two aphorisms of GM II contradictory.

My aim in this chapter is to argue that this contradiction is only apparent. What Nietzsche describes in GM II: 1 is the *capacity* that underwrites trust (permitted promising), and what he describes in GM II: 2 is the *exercise* of that capacity, realized to a high degree, as exemplified in the sovereign individual. Moreover, I argue that two different forms of trust correspond to each of these aphorisms, because the memory of the will can be realized in two different ways. A consequence of this is that one can have the "permission to promise" in a weaker and a stronger sense. The weaker sense is captured by conceiving of trust as an attitude of reliance from the participant stance (Holton 1994), and so involves regarding the agent as one who is a morally responsible agent, and not merely reliable. By trusting another in this way we might rely on any number of particular considerations, say the promiser's goodwill (Baier [1986]; Jones

[1996]), or an assessment of her character (Reginster [2017]), or these in addition to her competence in matters of judgment and her truthfulness more generally [Hieronymi 2008]). In any case, I suggest that, from Nietzsche's point of view, when understood in such terms trust of others requires only that they be able to keep their promises out of a sense of moral obligation, and therefore bottoms out in the agent's capacity to feel guilt. I call this *moral trust*. It is my focus in Section 3. There I argue that most humans are permitted to promise in this weaker sense on Nietzsche's analysis, i.e., we can be relied upon from the participant stance, because bad conscience and the ability to feel guilt ensured that most of humans are able to keep their sincere promises out of a sense of moral obligation.

However, Nietzsche's *ideal* responsible agent, the sovereign individual, keeps his promises for a different reason, or from different motives, which may indeed conflict with our moral obligations. He is "permitted to promise" in a stronger sense. As Nietzsche says, conscience is the sovereign individual's "dominant instinct" (GM II: 2); in him we witness conscience in its "highest, almost disconcerting form" (GM II: 3). I interpret this to mean that he not only has a memory of the will, or the *capacity* to extend practical commitment in the absence of external incentives, but that sovereign conscience is a *higher-order* commitment to preserving or maintaining one's values. As I will characterize it, sovereign conscience is a commitment to *integrity*, which bears an important relationship to Nietzsche's positive ideal of autonomy. This is why the sovereign individual represents an *ideal* of responsibility, more precisely, an ideal of what it means to be *trustworthy*. I will call this *sovereign trust*.

According to Nietzsche's ideal of sovereign trust, most of us are not trustworthy, at least not most of the time, *even* when we keep our promises from the participant

stance and do not violate the promisee's trust. To be clear, most humans are apt recipients of *moral trust* because we are internally motivated to keep our sincere promises out of a sense of moral obligation. But because most of us only pay lip service to our integrity, and because Nietzsche's ideal of autonomy is immensely difficult to achieve, few of us are ever *worthy* of making promises—at least according to the standards of sovereign trust.

So, by distinguishing moral trust from sovereign trust, I will argue that we can affirm both horns of the interpretive dilemma and dispel the scope problem. Because human beings in general are apt recipients of *moral trust*, our being "permitted to promise," understood in this weaker sense, is a status that distinguishes us from other types of animals. More precisely, we are "permitted to promise" in the sense that it would be appropriate to *accept* our promises, on the basis of moral trust. However, most humans are not *worthy* of trust, because we do not satisfy Nietzsche's ideal of trustworthiness, as exemplified in the sovereign individual, i.e., according to the standards of *sovereign trust*. In this stronger sense, most humans are *not* "permitted to promise," in the sense that we should not *make* promises. The reason for this is that keeping promises out of a sense of moral obligation turns out to be weak or deficient on Nietzsche's analysis, for reasons I will explain in Section 5.

I will pursue this solution by first critiquing some inadequate interpretations of Nietzschean trust. I call these "dispositionalist accounts" because these authors contend that having the permission to promise can be explained in terms of a reductive behavior analysis of trust.

II. Dispositionalist Accounts of Permitted Promising

Dispositionalist accounts of permitted promising are defend by Christa Davis Acampora (2006), Brian Leiter (2011), and Lawrence Hatab (2009). With respect to the scope issue, these scholars believe that most or all humans are permitted promisers, because the memory of the will confers nothing more than an ability to *remember* one's promises or debts (Acampora [2006: 148]; Leiter [2011: 108]; Leiter [2015: 179, 183]). As I argued in Chapter 2, these interpretations collapse the distinction between mere reliability and responsibility that Nietzsche invokes in GM II. As I will argue here, they also collapse the distinction between mere reliance and trust that he invokes, and therefore do not offer plausible interpretations of what it means to be "permitted to promise."

According to Brian Leiter, when Nietzsche opens GM II by asking what it would take to "breed an animal that *is permitted to promise*" (GM II: 1), all that he wants to explain is the "behavioral disposition of promise-making," or the creation of an animal that is "able to make and keep a promise" (2011: 106). Leiter goes on to state that we explain this disposition "not by appeal to [our] exercise of some capacity for autonomous choice and decision, but in terms of the causal mechanisms (e.g., breeding) acting upon [us] which yield certain steady behavioral dispositions" (Leiter 2011: 106). Acampora concurs. According to her, "Nietzsche is essentially asking: what sort of being, *what sort of animal*, must one become in order to be able to make promises?" (2006: 148). Does Nietzsche even care about *keeping* those promises? Her reading is less than clear on this point because she believes Nietzsche is trying to explain a rather basic a *Kraft*—a power or ability—how we became "capable of" making promises by overcoming forgetfulness (Acompora 2006: 149). Hatab endorses her analysis of conscience (2008: 171-2).

These deflationary interpretations of the sovereign individual are all based on offering a deflationary interpretation of sovereign *conscience*. As Leiter remarks,

> Surely it bears emphasizing that [the sovereign individual] is described as having one and only one skill: he can actually make and keep a promise! And why can he do that? Because he can remember that he made it, and his behavior is sufficiently regular and predictable, that others will actually act based on his promises. (2011: 108)

On Leiter's view, the memory of the will is nothing more than an ability to remember one's promises or debts (2011: 108; 2015: 179, 183). This is half the reason sovereign individuals are "permitted" to make promises; the other half consists simply in their being "regular" and "predictable" (GM II: 1, 2) in their behavior such that others will *accept* their promises. On this view, trust is ultimately grounded in a *behavioral disposition* to keep promises: *A is permitted to promise if A can remember the promise, and B accepts the promise on the basis of A's past (reliable) behavior*.

According to dispositionalist accounts, creditor-debtor relationships would be occasions in which a promiser was "permitted to promise" in this sense (see GM II: 5). Consider these as they exist today. If you want to take out a line of credit, or enter into some other kind of contractual relationship with a business (e.g., get a new cellphone plan), the first thing the creditor is likely to do is check your credit history. That report serves as a general indicator of how reliable you've been at repaying your debts in the past, and it is used by the creditor to decide whether you will pay your debts in the future, and to set the terms of the arrangement. Note, the creditor's willingness to enter this relationship is not determined on the basis of the sincerity of your intentions, and does not require an assessment of your goodwill, character, or your truthfulness. The decision is based solely on the basis of your past behavior when it comes to repaying your debts. Or consider creditor-debtor relationships as they existed in the past on

Nietzsche's analysis. The creditor's willingness to enter the relationship depends on the guarantee of satisfaction, in the exchange of the "directive and right to cruelty" (GM II: 5). The debtor agrees, when entering the relationship, that should he fail to repay he will be subject to "all manner of ignominy and torture" (GM II: 5). Again, the creditor does not care about whether you're sincere and truthful, or whether your will is good, or your character upstanding.

All of this points to the fact that these relationships are not founded on the basis of *trust*.[131] Intuitively, if the creditor *trusted* the debtor, there would be no need to threaten him with the loss of his limbs. His promise would be enough; his word alone would suffice. The reason for this is that, by making a promise of repayment, the debtor intends to produce a belief in the debtor, i.e., the belief that he will repay. However, by threatening him with the loss of his limbs, the creditor is doing nothing if not acknowledging that this belief has *not* been sufficiently produced by the debtor's promise. Similarly, the cell phone company wouldn't go through the hassle of checking your credit history and having you consent to a contract that includes all manner of provisions should you fail to repay your bill—if they *trusted* you to pay it. If such agreements were founded on the basis of trust, there would be no need for things like late penalties and collection agencies.

Creditor-debtor relationships are formed on the basis of mere *reliance* in the two aforementioned ways. They may be formed on the basis of what we can reasonably expect people to do when they promise, based on how they've acted in the past (the credit report), or they may be formed on the basis of what we can reasonably expect people to do when they promise, assuming we also provide them with some additional

[131] See Baier (1986).

and strong incentive to do it (the threatened loss of limbs, collection agencies). Both practices give the debtor the confidence to rely in the absence of trust.[132]

To rely on someone or something occurring is to "plan on it happening ... to work around the supposition that it will," according to Richard Holton (1994: 65). The same thing occurs when we trust—we rely on the person to do as they promised, by working their plans into our own. But reliance is *unlike* trust in a crucial respect. When we rely on another, we assume only that they are "regular" and "predictable" in keeping their promises (GM II: 2). We treat their behavior as evidence that their promise will be kept, and accept their promise on the basis of that evidence. But in that case we are only treating them like "a good thermometer" (2008: 222), to quote Pamela Hieronymi. We do not take them to be *responsible*. Put in terms of GM II's developmental analysis of conscience, when we rely on another all we do is *hold them to an expectation of conformity*. We expect the person to conform her behavior to some rule or convention ("I will not break my promise"), given our knowledge of her past behavior, or given the presence of incentives that would induce her to abide by the norms of promise-keeping. To rely on another therefore presumes *only* the existence of non-moral conscience, the memory of "I will nots" (GM II: 3).

However, as I argued in Chapter 2, Nietzsche does not believe that, to be responsible or permitted to promise, it is sufficient for one to simply be reliable or relied upon. First, as I argued there, he identifies reliability of behavior as a "presupposition" of the memory of the will, and so reliability cannot be the capacity conferred *by* the memory of the will. Secondly, the morality of custom was only a "means" to producing

[132] Note, Nietzsche claims the debtor must "instill" or "inspire" [*einzuflössen*] *Vertrauen* [trust] in the creditor, because it is something that must be artificially manufactured by the offer of collateral.

the sovereign individual (GM II: 2), and it made us reliable, so being reliable and being *responsible* cannot amount to the same thing.[133] Conclusive evidence of this is provided in GM II: 3. There Nietzsche explains the origin of non-moral conscience as a memory of "I will nots," but he *contrasts* this with sovereign conscience:

> His conscience? ... One can guess in advance that the concept 'conscience,' which we encounter here in its highest, almost disconcerting form, already has behind it a long history and metamorphosis. To be permitted to vouch for oneself ... that is, as noted, a ripe fruit, but also a *late* fruit:—how long this fruit had to hang on the tree harsh and sour!"

The "harsh and sour" fruit is the non-moral conscience. The "ripe" and "*late*" fruit is the "memory of the will," specifically as instantiated in the sovereign individual. The former is preliminary to the latter, and the two therefore cannot be the same.

In Chapter 2 I argued that difference between these two types of "memory" consists in the *motivation* associated with each. The memory of the will is "an active no-longer-wanting-to-get-rid-of," and "thus by no means simply a passive no-longer-being-able-to-get-rid-of" (GM II: 1). Non-moral conscience is a memory of the expectations of *others*, of what they might do to me if I don't conform my behavior to generally agreed upon standards or norms. In this case I keep my promise because I cannot rid myself of the thought of what might happen to me if I don't (hence its being "passive"). A memory of the *will*, by contrast, is an ability to *sustain* or *perpetuate* the motivation to keep one's promises (hence its being "active").[134] As I will explain later, the memory of the will assumes that the agent is *internally motivated* to keep her promises, or to honor her values or commitments more broadly, which might be achieved in two different ways. The agent might regard these as "personal obligations" (GM II: 8) that bear on her sense

[133] See Chapter 2, §2.
[134] See Reginster (2011: 58).

of worth, and so keep them out of a sense of moral obligation attached to guilt (§3.1), or she may keep them out of a motive of conscience or integrity, which I argue is the motive involved in sovereign promising (§4.2).[135] For now, the lesson to draw is that the type of motivation characteristic of the "memory of the will" is *not* contingent on others holding me to an expectation of conformity.

According to Holton, internal motivation is one feature that distinguishes trust from mere reliance. As he says, "To [trust a person] is not just to rely on a certain state of affairs happening: the state of affairs in which they do that thing. Rather, it is to rely on them doing it from a motivation that stems in some way from them" (Holton 1994: 65-6). To illustrate the importance of internal motivation for promise-keeping, I will now consider one possible reading of the "frail dogs" Nietzsche derides. These individuals make promises "although they are not permitted to do so" because they are "lesser-willed" and "unreliable" (GM II: 2). Nietzsche's point is that this person cannot be trusted because his will is too weak for him to bind himself to a course of action. We will consider this *weak-willed* frail dog in Section 4. Another, more extreme kind of frail dog is more germane to dispositionalist accounts. On this stronger reading the frail dog cannot be trusted because he is not internally motivated *at all* to keep his promises. That is, he lacks a "memory of the will" entirely, at least in matters of promise-making and promise-keeping.[136] Call this the *unmotivated* frail dog.

---

[135] According to Clark and Dudrick (2012), all acts of will engage an agent's internal motivational system on Nietzsche's analysis, because willing paradigmatically involves overcoming sources of resistance in an effort to realize the agent *values*. As Clark and Dudrick suggest, it would be more accurate to characterize Nietzschean willings as acts of *will power*. The analysis of conscience as a memory of the will that I provide here will lend support to their analysis, especially as exemplified in the sovereign individual.

[136] This stronger reading is more germane to dispositionalists because they interpret the memory of the will as the "passive" form of memory. It is especially germane to Leiter's interpretation, because he takes the will to be epiphenomenal on Nietzsche's analysis. As

An unmotivated frail dog is not permitted to promise, but we are nonetheless taking a number of things for granted about him relevant to promising. We are taking it for granted that he is a socialized human being, one who has a familiarity with and in general conforms his behavior to, obligations. He therefore has the ability to *make* promises.[137] He also has the ability to *keep* them, at least when promises are made in a scenario or an environment that provides him with strong enough incentives to do so. That is, he can keep them when fear, the prospect of reward, or the damage to his reputation prove strong enough for him to keep his promises. However, he is not permitted to promise, or worthy of trust, for this reason. He does not keep his promises out a sense of personal or moral obligation, or a sense of integrity, as one who possesses a memory of the will does. Instead, he needs to be monitored and manipulated in various ways to ensure that he keeps his promises. Or he may keep his promises out of an entrenched sense of habit, because of the "herd instinct of obedience" (BGE 199), but in that case we have no assurance that he would keep his promise if provided with a

---

evidence of this he often quotes *Antichrist* 14: "Formerly man was given a 'free will' as his dowry from a higher order: today we have taken away his will altogether, in the sense that we no longer admit of the will as a faculty ... the will no longer 'acts' or 'moves'"(A 14). It does seem as if Nietzsche is denying the existence of the will in this passage, but Leiter never tells us how this passage is to be made consistent with GM II: 1 and other passages where Nietzsche clearly assumes the will/willing is causally efficacious (see BGE 21, TI V: 2, TI VIII: 6). The passage suggests a ready answer, though not one in support of epiphenomenalism.

What Nietzsche rejects here and elsewhere is *agent-causal* free will (see GM I: 13, BGE 21), or the "faculty of will" understood to be a kind of agent-cause *distinct from* an event-cause (hence its being a gift from a "higher order"). The human will, like every other kind of cause, is an event-cause, because every act of will is determined by our unconscious drives. As Nietzsche says, "The old word 'will' now serves only to denote a resultant, a kind of individual reaction, which follows necessarily upon a number of partly contradictory, partly harmonious stimuli" (A 14). These "stimuli" are our unconscious drives. What he says here is perfectly compatible with BGE 19 where he also denies—not the existence of will or our ability to effect change by willing—but only the proposition that willing is "simple," as Schopenhauer contended.

[137] "[M]an himself must first of all have become *calculable, regular necessary*, in his own image of himself as well, in order to be able to vouch for himself *as future*, as one who promises does!" (GM II: 1). As I argued in Chapter 3, because making a promise involves communicating an intention to undertake an obligation, it presumes that the promiser has a general familiarity with obligations and a sense of oneself as a reliable, i.e., the rudimentary conscience as a memory of "I will nots" (GM II: 3).

strong enough countervailing incentive to renege on it. In short, we can *rely* on him to keep his promise in environments where incentives exist.

However, we cannot *trust* him, according to Holton's first criterion of trust.[138] The character Littlefinger from *Game of Thrones* illustrates why this is so. Littlefinger is motivated only by power, political ambition, and winning the affection of two women he is obsessed with but who do not reciprocate his affection. He has no moral scruples whatsoever and will do whatever it takes to achieve these aims. He is a man utterly devoid of guilt and moral conscience, though he certainly is "regular" and "predictable" in his behavior. If Littlefinger makes a promise, which for our purposes we will assume he only does to fulfill one of these aims, you can be sure that he will keep it. More accurately, Littlefinger keeps his promises *so long as* they are necessary to achieve any of these aims. Littlefinger is not capricious; he is *very* reliable. We know what to *expect* from Littlefinger, and so we can "work him into our plans."

According to proponents of the dispositionalist account, Littlefinger is someone you can trust so long as you have leverage over him, so long as the promise he made to you presents the necessary, best, or most efficient means of achieving these broader aims, because agents are permitted to promise so long as they possess a (passive) memory of their debts, and reliably conform their behavior to them. *Why* they do so is immaterial, but Littlefinger illustrates that the reason or motive is *not* irrelevant when what we do is trust and not merely rely. The problem Littlefinger presents for trust is that we know he is not motivated to keep his promises *as such*, that by making a promise

[138] Holton admits, "Exactly what does count as an internally driven motivation is not clear to me," but he rules out fear as a possible candidate (1994: 66). I will provide two possible answers. When it comes to moral trust, one's motivation is internal if violating it would elicit *guilt*, or if the agent is capable of keeping it out of a sense of personal or moral obligation. The sovereign individual, by contrast, keeps his promises out of a motive of conscience or *integrity*.

now he will not see that as a decisive or even weighty reason to so act *later*. He will do so only if that promise proves to be necessary, and perhaps the best or most efficient means, of achieving his other aims. So we are always left wondering if, in the interim, some other opportunity has come along that will cause Littlefinger to betray us.[139] If Nietzsche thinks that those who are internally motivated and yet too weak-willed are not permitted to promise, then it cannot be the case that those like Littlefinger, who are not internally motivated *at all* to keep their promises, are permitted to do so, either.[140]

As Littlefinger illustrates, *one* way that the "memory of the will" can prove defunct or inoperative is that one can "forget" that a promise made *then* is a decisive reason to so act *now*, because having been presented with a better opportunity in the meantime, the agent now lacks entirely the motivation to fulfill his promise. We might say that such agents do not feel the "force" of promises, or do not recognize them as *binding* commitments. They see their promise as binding only if it is *we* who provide them with that reason. Whether weak-willed or simply unmotivated, frail dogs are not permitted to promise on Nietzsche's account, and so the dispositionalist account of Nietzschean trust is inadequate. It cannot be that *A is permitted to promise if A can*

---

[139] You might think Littlefinger is not reliable for this reason, because we can never be sure that his promise is still being worked into *his* plans, and without knowledge *his* plans we cannot *confidently* rely. This is true, but whereas as a kind of confidence, optimism, or hope is arguably a condition of moral trust (see Section 4), these are *not* a condition of mere reliance. As I argue in Section 3, you might still rely on Littlefinger so long as you possess the *leverage* to ensure that he keeps his promise, so long as you believe can provide him with an overriding incentive to do so. If you take yourself to possess this kind of power, as a king arguably would, Littlefinger's failure to keep his promise amounts to a kind of non-moral betrayal.

[140] As Reginster observes, "Note that we do not call 'responsible' just anyone on whom we can depend to do what he has promised to do. For instance, we will not judge trustworthy or 'responsible' an agent who keeps his promises only because he fears the unpleasant consequences of breaking them or desires the rewards expected from keeping them. One plausible motivation for our attitude in this case is the recognition that the agent does not care about keeping his promises *as such*, and would break them the moment the unpleasant consequences of so doing would be either avoidable or outweighed by the pleasures for the sake of which the promises would be broken" (2011: 71).

remember the promise, and B accepts the promise on the basis of A's past (reliable) behavior.

III. Moral Trust

The second feature that distinguishes trust from mere reliance, according to Holton, is the kind of *regard* we extend to those whom we trust. As he explains:

> I think that the difference between trust and reliance is that trust involves something like a participant stance towards the person you are trusting. When you trust someone to do something, you rely on them to do it, and you regard that reliance in a certain way: You have a readiness to feel betrayal should it be disappointed, and gratitude should it be upheld. In short, you take a stance of trust towards the person on whom you rely (Holton 1994: 67).

According to Holton, trust is an attitude of reliance, and thus shares in common with reliance more generally the notion of anticipating or expecting performance, as noted previously. What distinguishes trust from mere reliance is the "stance" we take toward those whom we trust, and this stance presumes the first criterion of trust, that those whom we trust are internally motivated to fulfill their promises, or committed to keeping their promises *as such*.

According to Holton, this stance is evident by our proneness to experience moral reactive attitudes. Consider Littlefinger. We do not trust him because we know he will keep his promise *only if* doing so fits into his broader interests or aims, or if we have the leverage to ensure that he does. Knowing this about him, we are not deceived by his motives and intentions. Specifically, we would not expect the failure to keep his promise to elicit *guilt*, the "feeling of personal obligation" (GM II: 8), and so we should not rely on him from the participant stance, since that stance depends on an agent's ability to feel guilt. So, if Littlefinger reneges on his promise we should not feel resentment of the kind Strawson talks about, because by accepting his promise we should not have assumed that

Littlefinger has an ounce of *goodwill*. Similarly, Littlefinger would not be deserving of *gratitude*, because if he keeps his promise he did not do so out of a concern for our well-being, or to protect the vulnerability we entrust to others when we accept their promises on the basis of trust. We might be *relieved*, but we should be grateful for nothing. All of this points to the fact that we would not and should not have regarded him as a *responsible agent*. We should not have assumed he will *hold himself* to a *moral* expectation to fulfill his promise, because we should not have assumed he would regard it as a personal obligation.

I do think Nietzsche would disagree with Holton's analysis in one important respect, though. Littlefinger *can* betray us, and the reason for this is that betrayal is not only a failure to care for an entrusted vulnerability. In fact, this is precisely the sense in which Littlefinger *cannot* betray us, because it would be irrational, for those of us who know better, to expect him to keep his promises out of a concern for our well-being. Yet there is nothing irrational in the thought that one might nonetheless want *revenge* against him. Revenge is a "reactive affect" on Nietzsche's broader view of these emotions (see GM II: 11).[141] Like anger, it is a response to suffering some injury, and wanting revenge, like experiencing anger, is perfectly natural. It's just that wanting and seeking revenge is *immoral*, at least in terms of what morality requires of us today. (For the ancient Greeks, on the other hand, it was "sweeter than honey.") Today we can demand "justice," which is milder, controlled, and socially sanctioned revenge, but it would be

[141] I suspect revenge would not qualify as a "reactive attitude" in Strawson's sense, because these all depend on the demand or expectation for goodwill or regard (1962: 76, 90), whereas revenge is generally understood to be a response to offending one's power or standing. Perhaps Strawson's idea of "regard" includes the latter considerations, but I see little evidence of this. Also, one might think that the agent who experiences revenge ought to recognize that it is *immoral*, because desiring it also violates the demand for good will or regard, and ought to give it up. However, Strawson acknowledges that resentment has this character as well (1962: 90), and he doesn't think it problematic for that reason.

wrong of us to want pure revenge, which is hot, personal, and in principle knows no bounds (see GM II: 9-11). This shows a *second* way Littlefinger might betray us, by betraying our sense of fairness or justice, which I argued in Chapter 4 (§2) depends only on considerations surrounding objective guilt. Littlefinger might betray one in this way by making us believe he would abide by the norms of promise-keeping he is well aware of, which every sincere promiser upholds, and which he exploits for his personal gain when he lies and cheats to achieve his personal ends.

Yet there is a *third* way that Littlefinger might betray us, even if he does not betray our sense of justice or fairness. Again, Littlefinger might be able to take advantage of unsuspecting and gullible persons by making them think he has a sense of justice or fairness, but we know better. And yet we might still want revenge when he breaks his promise. Why? Because by breaking his promise he can still offend my sense of *entitlement* as one who expected and demanded performance from him, because he made a promise in which said performance was purportedly assured. To not follow through is thus a kind of betrayal, not of my trust, because I am not assuming there is an ounce of goodwill or justice in him, but of my *power*. This would be like Littlefinger breaking his promise to a king, for instance. The king expects Littlefinger to keep his promise because he takes himself to possess the *leverage* to ensure that Littlefinger does, which Littlefinger offends or disregards by instead assuming the king to be a weak kind of person who can be disrespected without repercussion. If this is right, there exists a kind of *non-moral* betrayal that does not depend on the demand for goodwill, regard, or a sense of justice or fairness.[142]

[142] I take it to be obvious that revenge is problematic from a moral perspective, but is it unfair in this case? I suspect Nietzsche would say no. Littlefinger is no idiot. He is conniving and

That said, I take it Nietzsche would agree with Holton's main point that we do not trust Littlefinger because we would not regard him as a *responsible agent*. That is, we may rely on him, but not from the participant stance. Let me clarify this by explaining why, by contrast, most of us *are* "permitted to promise" in the weaker sense that Holton is describing. We are in general apt recipients of *moral trust*, because we are capable of being internally motivated to keep our promises, and generally *do* keep our sincere promises out of a sense of moral obligation. It is the ability to feel guilt, and the kind of regard of it makes possible, which demarcates human beings as permitted promisers from other animals in the natural order of things.

3.1 Moral trust and Guilt

Holton provides a general view of trust as reliance from the participant stance, but that attitude itself might track any number of more specific considerations.[143] By trusting another we might rely, among other possible things, on the promiser's goodwill (see Baier [1986]; Jones [1996]), or an assessment of her character (see Reginster [2017]), or an assessment of her competence and truthfulness in matters of judgment more generally (see Hieronymi [2008]). The important thing to note about trust, whatever specific consideration(s) it tracks, is that "*trust links reliance with responsibility*," as Margaret Urban Walker says (2006: 80). "In trusting one has *normative expectations* of others," she explains, "expectations of others that they will do what they should and hence that we are entitled to hold them to it, if only in the form of rebuking and demanding feelings" (Walker 2006: 80).

---

extremely intelligent. He knew that betraying a king would in result in, say, death and thus frustrate his personal aims, and yet that is a risk he took voluntarily because he took himself to be so clever.

[143] He critiques Baier's suggestion that trust always depends on another's goodwill toward one (Holton 1994: 65), though he presumably would not disagree that it often does. If this is right, Baier identifies a plausible reason underwriting moral trust, just not the *only* one.

By "should" Walker does not mean the prudential hypothetical "should" of instrumental reason, nor the non-hypothetical "should" of non-moral obligation, the latter which I have argued rests on expectations of conformity. She means "should" in the sense that Nietzsche means "should" in GM II: 15, when he says for a long time the thought "I should not have done that" never occurred to criminals. That is, she and Nietzsche both have in mind the moral-categorical "should," a thought which never occurs to people like Littlefinger because they are incapable of experiencing "pang[s] of conscience" (GM II: 14). On Walker's account, and on all of the above analyses trust, trusting another depends in some way or another on holding the promiser to a *moral expectation* to do what is promised, because by trusting we rely from the participant stance, presuming that the promiser is a morally responsible agent. Accordingly, it is appropriate to experience moral reactive attitudes like resentment, betrayal, and gratitude, because our proneness to these is constitutive of the kind of regard trust depends upon. According to Nietzsche's analysis of responsibility in GM II, trust would not be possible if agents were incapable of undergoing episodes of bad conscience. More precisely, the *regard* we extend to others when we trust them, as well as the kind of *internal motivation* characteristic of the memory of will, are both predicated on the agent's ability to feel guilt.

Consider, first, how guilt makes the stance or regard of trust possible. A friend asks to borrow some money and you know him to be sincere in his intention to repay it and upright in his character. You're confident that he would not bilk you out of your money, and so you loan him the money relying only nothing more than his goodwill. You do not setup a payment plan, specifying that you will perform monthly audits to ensure the payments are made in a timely manner, or ask him to offer some collateral in case he

doesn't repay, and you don't inform third-parties of the loan (e.g., friends you share) so that they will encourage him to repay. No, you accept his promise on the basis of his goodwill alone. Unlike creditor-debtor relationships, you become *vulnerable* by accepting his promise in this way, a vulnerability that is entrusted with the promiser to care for and weight appropriately. Apart from expected performance—apart from you *relying* on your friend to repay—you take him to be *responsible* for protecting that vulnerability. It would be incoherent to task another with the protection of this vulnerability if they were incapable of feeling guilt, on Nietzsche's analysis.

Guilt is the self-reactive feeling of responsibility; it manifests as a desire to punish oneself for violating "personal obligations" that bear on our sense of worth as persons. By being attached to guilt, the obligations we voluntarily undertake are thereby given the appropriate kind of weight and priority when deciding how to act, in reference to the specific kinds of considerations that trust is tracking (e.g., my goodwill). In the above example, this corresponds to the relatively stringent obligation to repay a friend, but this is true of moral obligation more generally as well. When Littlefinger makes a promise, he is incapable of seeing the promise *itself* as providing any reason to restrict future possibilities of action. We, on the other hand, *do* recognize that reason, because we recognize and appreciate the moral significance of making and keeping promises. If the choice is between betraying a friend by taking the money and ghosting, or repaying a friend who did you a solid, the fact that we would experience guilt for the former choice provides a compelling pro tanto reason not to do it. In general guilt allows us to see and appreciate the moral significance of our actions in this way, because guilt enables us to *relate* to obligations in a novel way.

Littlefinger also lacks the ability to hold himself responsible, because he lacks this novel comportment to obligation. We therefore should not *regard* him as a morally responsible agent, if and when we do accept his promises. It is therefore inappropriate to blame and resent him, if these are intended to elicit guilt in an effort to induce an apology, to get Littlefinger to show contrition, or make amends out of a sense of remorse. He is quite immune to such overtures, and so if Littlefinger betrays us and we demand satisfaction, we will have to achieve it in other ways (i.e., revenge).[144] The friend case is different. There we are to imagine our blame and resentment has some effect on him, because he made the promise sincerely and had every intention of keeping it. If blame is warranted and intended to elicit guilt (e.g., he invested the money in a successful startup but failed to pay the taxes), it is appropriate for him to feel guilty and to express remorse or contrition as a first step toward making amends. His feeling guilt is in this way an expression of his *taking* responsibility or accountability, which I have argued moral blame depends upon.

Finally, the kind of *internal* motivation that is constitutive of the memory of the will and of trust also assumes that we have the ability to feel guilt. The memory of the will is an ability to sustain practical commitment in the absence of external incentives, and the advent of bad conscience made this possible insofar as it gave "depth" and "breadth" to the human soul (GM II: 16). As I described in Chapter 4 (§3), this is essentially a matter of our developing higher-order desires, values, and commitments, and our acquiring the ability to hold ourselves to expectations corresponding to these, which created the possibility of conflict with our first-order desires and drives. Those

[144] A possibility I would be remiss to mention is that we simply let Littlefinger go unpunished, by showing him mercy. Nietzsche raises this possibility and by all appearances seems to advocate it, but it "remains the privilege of the most powerful, better still, his beyond-the-law" (GM II: 10). I take it most of us are not capable of being merciful for this reason.

higher-order values and commitments that would elicit guilt are "personal obligations" and integral to our practical identity. We might imagine a structural or hierarchical view of the self—on Nietzsche's view, a "political order" of the drives, affects, and values (BGE 6, 9, 12; see Clark and Dudrick 2012: Chapter 7). The kind of motivation constitutive of the memory of the will is premised on our having such a "political order" or practical identity. Without it the ability to will in the absence of external incentives would be impossible. The reason or this that, to be "internal" rather than "external," the agent must see the motive as in some sense stemming from her, as opposed to forces outside of her (even if those forces are "internal" to her body). As Frankfurt says, she must "identify" with the internal motives, taking them to have a kind of privileged standing or authority within the self's constitution. Without such identification, it would be incoherent to suppose an agent could sustain practical commitment through an act of will alone, because all commitment would in that case simply be a matter of incentives.

The ability to feel guilt thus makes possible both the kind of *regard* and the type of *motivation* constitutive of moral trust. It ensures that we give our sincere promises the right kind of deliberative priority to be tasked with the vulnerability entrusted to us, given that a broken promise will impugn our good will, our character, or our standing as a competent and truthful person. Guilt imbues promising with moral significance in this way, enabling us to see that making a promise is *itself* a reason to keep it, because by making a promise I have placed myself under a personal or moral obligation. Guilt make us *responsible for* our promises and not merely reliable at keeping them, because we *hold ourselves* to an expectation to keep them out of a sense of moral obligation. And so we keep our promises not merely because others hold us to an expectation of conformity. In short, guilt enables the kind of regard constitutive of the attitude of moral trust.

3.2 The Scope Problem

When Nietzsche introduces the conscience as a "memory of the will" in GM II: 1, what he describes there is the *capacity* underwriting permitted promising, or trust. He assumes that most humans are permitted to promise in the sense that they possess this capacity, because the problem of "forgetfulness" which the memory of the will evolved to address has been "solved to a high degree" (GM II: 1). This "problem," as we have seen, is not solved by creating a "passive" memory of one's obligations so that one will not forget them; it requires the creation of an "active" memory that enables the agent to sustain the motivation to do so in the absence of external incentives. When Nietzsche characterizes conscience in these terms, he takes it to be a faculty directly relevant to the will, and to the activity of binding oneself to a course of action. Conscience is also, for this reason, directly relevant to making and keeping promises, and to trust, insofar as promising is an act of binding oneself. Most humans possess this ability on Nietzsche's analysis. That is, most of us are capable of binding ourselves by making promises to others and seeing them through, because in general we can and do keep our promises out of a sense of moral obligation. As I have just argued, doing so presumes that we are capable of feeling guilt, because guilt is constitutive of the attitude of moral trust.

The ability to feel guilt is an ability Nietzsche takes most humans to possess, as I argued in Chapter 4. If he did not think this, bad conscience could not have represented humanity's "forceful separation from [our] animal past" (GM II: 16). The ability to feel guilt is therefore a capacity that distinguishes human beings *in general* from other creatures in the natural order of things. For Nietzsche, the boundaries of *exemption* track the boundaries of guilt. Therefore, the animal that is "permitted to promise" also corresponds to the boundaries of guilt. Specifically, most humans are "permitted to

promise" in the weaker sense that they are apt recipients of moral trust, because they are capable of and generally do keep their sincere promises out of a sense of moral obligation. But this does *not* qualify one as a "sovereign individual."

An interesting but often unnoted fact is that when Nietzsche first introduces the idea of permitted promising, he does so by referring to the human being as the *animal* [*das Tier*] who is permitted to promise (GM II: 1). The relevant contrast class is other animals, not other human beings. The scope of this designation changes when he shifts to a discussion of the sovereign individual, where he is identified as the *human being* [*Der Mensch*] who is permitted to promise (GM II: 2). This implies, strangely enough, that the standards of trust change along with the shift in contrast class, depending on whether we are relating humans to animals or humans to other humans. Below I will try to lend credence to this strange idea, by connecting Nietzsche's ideal of sovereignty and autonomy to his critique of moral obligation. This has been a neglected topic thus far, and so it is about time we said something about it.

IV. Sovereign Trust

Nietzsche would be the first to remind us that it is one thing for trust to be *given*, and quite another for it to be *earned*. Trust may be given for all sorts of reasons, indeed reasons that are quite at odds with the idea that it might first need to be earned. This is particularly clear from the way Karen Jones, Victoria McGeer, and Margaret Urban Walker talk about trust. According to Jones,

> Trust is an attitude of optimism that the goodwill and competence of another will extend to cover the domain of our interaction with her, together with the expectation that the one trusted will be directly and favorably moved by the thought that we are counting on her (1996: 4).

According to McGeer, trust is underwritten by an attitude of hope:

> [Trust] is an attitude that both empowers *us* in our trust-making it possible for us to think and act in trustful ways—and empowers *them* through our trust, by stimulating their agential capacities to think and act in trust-responsive ways. Since this state of mind is forward-looking, anticipating the transformative effects of extending our trust, I think it is most aptly characterized as an attitude or condition of hope. (2008: 242)

As we saw above, Walker believes that "*trust links reliance with responsibility*" (2006: 80) through the medium of normative expectations. Those expectations *trusting* when they

> embody a *hopeful* attitude, one that includes a belief in the possibility of responsiveness, and a desire for responsiveness, alongside a demand for responsiveness even in the absence of reasons for optimism. In this kind of case, our attitude of normatively expecting something of others (or ourselves) "pulls for" correct behavior in the face of uncertainty, and possibly in the face of evidence to the contrary. (2006: 69).

Walker and McGeer disagree with Jones about the affect underwriting trust. It is hope rather than optimism, they contend, because in many cases of trust we lack the reasons or confidence to trust *optimistically*, and in such cases we instead trust *hopefully*.

The difference is immaterial with respect to Nietzsche's concerns, and I have no reason to suspect he would object to their characterization of moral trust in many instances. When the mechanic informs me that my car needs new brake pads, or the HVAC repair man says the broken AC unit needs a new condenser, it is natural to hopefully or optimistically trust that they are telling the truth and can competently do the job. One *could* get a second opinion, but what if we are in the midst of a heat wave or driving halfway across the country on a family vacation? *Trust but verify* is not a viable option in such times, and so we seem to be left with two options: skeptical mere reliance or optimistic and/or hopeful moral trust. Nietzsche need not disagree.

It also bears mentioning that Jones, McGeer, and Walker all care about the rationality of trust, and so do not advocate trusting just anyone, even if they believe doing so paradigmatically involves a healthy degree of optimism or hope. (I take it none of them would advise us to trust Littlefinger.) I mention their accounts here because they make particularly evident the contrast we began by noting and which is actually fundamental, I take it, to Nietzsche's interest in permitted promising. It is possible and indeed rational to trust others on their accounts in the absence of considerations that would first show the agent to be *deserving* of trust. Indeed, I will argue this is true of moral trust more broadly on Nietzsche's analysis. By making a promise, the promiser takes himself to have the kind of foresight, self-command, and fortitude to *bind himself* to a future course of action, especially when and if significant obstacles lay in his path. But moral trust both obscures this fact and does not actually seem to require this of us. As we saw above, in general it demands only that we be capable of keeping our promises out of a sense of moral obligation, and this obscures the fact of whether the agent is capable of *binding himself*, because it shows only that he succeeds in binding himself *through another*. Promising on the basis of moral trust therefore exhibits a kind of weakness or deficiency, specifically one related to autonomy, as I will argue here.

4.1 Moral Trust is Tenuous

To appreciate the force of Nietzsche's argument, we first need to recognize that moral trust is tenuous in all sorts of ways that do not necessarily register as violations of trust by the promisee. When we rely from the participant stance, we assume that the promiser will be favorably moved *by* our reliance, in a way that is consistent with honoring the vulnerability entrusted to them. Often that is not the case. I invite the reader to consider the many reasons you might keep your promises, and which of them generally wins out.

I know in my case I *sometimes* keep my promises out of a sense of moral obligation, but I also and more generally do so because I'm worried about things like my reputation, getting something else I want (a reciprocal agreement), or would like to avoid being inconvenienced by the additional and unnecessary hassle that tends to accompany a broken promise. None of these motivations are consistent with moral trust.

Imagine that you ask the HVAC repairman to return tomorrow at the earliest possible time to replace the condenser (it is very hot after all). Seeing your desperation, he agrees. He promises to show up between 8-10 AM. We would *like* to think that shows up on time because we are relying on his goodwill, or because he takes himself to be an honest person. But the *real* reason he shows up is that he wants to get paid for the job, and he doesn't want you to leave a negative review of his business on Yelp. If so, he wasn't actually worthy of your trust. But he did keep his promise and so it did not register as betrayal, either. You trusted hopefully or optimistically, and he did not let you down. This is one reason moral trust is tenuous.

Or consider a case where money is not involved. You ask a coworker, Peggy, to attend your daughter's first birthday party, and she promises to show up. Again, you'd like to think Peggy shows up because she cares about you or your daughter, and this is true, or else she would not have agreed to come. But as often happens, a more attractive opportunity presented itself in the interim (some friends of hers are going to the beach). She shows up to the party anyway, but is now going to have to leave early. No big deal; this sort of thing happens all the time. On the face of it this is not a violation of moral trust and would seem to exemplify it, because Peggy *appears* to be showing proper regard for you and your daughter by showing up. However, imagine that she was deliberating the night before whether to attend the party at all. It would be far more

convenient, she realizes, to skip the party altogether and go straight to beach (traffic is always lousy on Saturdays). She is about to send a text declining the invitation when it suddenly occurs to her she will catch flack about it in the office on Monday. We might imagine her instead texting her beach-going friends, "I have to make an *appearance* at this birthday party I promised to attend, but I promise to meet you at the beach shortly after." If this is indeed her rationale and she kept her promise for this reason, she showed *some* degree of regard for you and your daughter. Indeed, her doing so was precisely enough to generate the initial conflict about what to do. It's just that what she felt she ought to do wasn't strong enough to get her to do it, until she realized she also would catch flack for it on Monday. So you might think she did not show the *proper degree* of regard. The reason for this is that her motive, though it was partially moral, was not *pure*; it was contaminated by a heavy admixture self-interest. This is another way moral trust might prove tenuous.

Like the HVAC repair man, it seems Peggy wasn't worthy of your trust, but we were none the wiser, so her action did not register as betrayal or a violation of trust. (No one has shattered your optimism or hope, or given you any reason to think them untrustworthy.) Perhaps her intentions will come out later (you noticed she looked disinterested at the party, and you ask her why on Monday), but perhaps not. The point is: moral trust functions perfectly fine in the *absence* of reasons to think agents are unworthy of it, and yet often agents (arguably) *are* unworthy of it. Nietzsche is especially sensitive to these sorts of concerns because inner opacity is a *big* problem on his view (see Riccardi [2015]). Essentially, the problem of inner opacity is that we lack reliable knowledge of, and introspective access to, the *real* causes of our actions—our subconscious drives on Nietzsche's view—and we instead rationalize our actions after the

fact, and make moral judgments on the basis of consciously felt motives that *distort* of these underlying drives.[145] So, not only is my knowledge of *my* intentions and motives incomplete, but my knowledge of *others'* intentions and motives is also incomplete, and yet moral trust depends on making such inferences. And so it rests on tenuous grounds.

However, the problem of inner opacity is not Nietzsche's main critique of moral trust. Let's consider a modification of the Peggy case that really goes to the heart of the matter. Same scenario, you ask a coworker to attend you daughter's party, but this time she is a friend and not just an acquaintance. Call her Claire. She shows up just the same and is going to leave early, but this time she is moved solely by your entrusting her with a vulnerability. In other words, her rationale is *not* "I have this other obligation I need to make an appearance at, or else it will be inconvenient for me," but rather "I promised a friend I'd attend her daughter's first birthday party, so I'll be a bit late to the beach." I take it her rationale exemplifies moral trust because she is moved for the right (moral) reasons. Nietzsche nonetheless wants to argue that moral trust in this case rests on a tenuous basis, because this form of promising is weak or deficient. In fact, he would argue that Peggy is in this case promising like a *weak-willed* frail dog. Why?

The first thing to note is that, though Claire keeps her promise for reasons that are consistent with moral trust, she *is* showing up halfheartedly. All we mean by this is that she is in a state of conflict, that her *will* is compromised. She wants to do two things, both are important to her, but she cannot do both at the same time, and doing the one will impair her ability to do the other. If she leaves the party too early, she'll miss

[145] Judgments distort primarily because our motives must be interpreted and conceptualized through consciousness and language, both of which developed due to the need to communicate with others, not to understand the innerworkings of our psychology (see GS 354). Rationalizing our actions and judging those of others is like using a magnifying glass to assess our reasons when what we need is a microscope.

out on the birthday cake or the opening of the presents. If she gets to the beach too late, the sun will set and she'll miss out on companionship with her other friends. If she maneuvers adeptly, she can keep both promises in a way that doesn't violate moral trust in either direction, in a way that would not open her up to resentment. Say she stays for the presents but not the cake. She tells you that she regrets having to go, but she really must leave. She *does* regret it—you can tell. You're grateful she came and thank her for coming—no ill will. She gets to the beach late in the afternoon but still in time to take a dip in the ocean before the sun sets. Her friends playfully rib her about showing up late, but in a way that makes it clear they are grateful she came. Again, no ill will.

I take it that Nietzsche thinks this kind of promising is pervasive. It is certainly familiar enough to me. But why is it *weak*? What is *deficient* about it? The problem is not just that she keeps both promises halfheartedly, or that her will is compromised. That we might think is just a basic fact about the human condition. We can't always do what we most want to do, and sometimes we must do things that we also recognize are important or obligatory. This conflict creates no special problem for moral trust so long as agents are favorably and adequately moved by the vulnerability entrusted to them, as I take it Claire was. As finite creatures we are also limited in our ability to divine the future, and so what happens in between the time we make a promise and execute it may be complicated by a myriad of factors (the most important one perhaps being time itself). But these factors do not necessarily show, as in the case just considered, that we are incapable of keeping our promises from a motive consistent with moral trust. The problem is a more fundamental and basic one concerning Claire's ability to *bind herself*.

Promising within moral trust works something like this. I make a promise to you and thereby place myself under a moral obligation to perform it. By accepting this

promise you then acquire a kind of entitlement or authority over me, to expect or demand that I do as I promised. Unlike simply forming an intention or deciding, in which case I can *unbind* myself at any moment, by making a promise to you I am committed to acting in the future as I said I would. Note, however, that I become bound precisely by binding myself *through you*. Once you accept my promise, it is no longer up to *me* whether I will do what I said I would, but up to *you*. So the reason that the future does not remain completely open to me is *your expectation*, and the way that your reliance registers with my internal motivational system (I know you're relying, say, on my goodwill, which motivates me to keep my promise, because I care about you). In short, at least part of the reason I keep my promise is that, if I don't, I will be committing a moral wrong, and I want to avoid committing this moral wrong.

This is how Claire acts. She moves adeptly in a way that pleases everybody, motivated by the thought that disregard in either direction will result in resentment among friends that she does not want to incur. We need not portray her as some kind of cold Kantian moralizer who reduces friendly relations to abstract and distant relations of duty—in moving in a way that she pleases her friends she also pleases herself. She kept her promise not just because she regarded it as a "personal obligation," but also for the more stringent reasons demanded of a friend. Yet she did keep her promise out of a sense of moral obligation. And when we keep our promises out of a sense of anticipatory guilt in this way, Nietzsche thinks our promise is weak or deficient.

To be clear, the problem is not Claire needs external incentives to keep her promises. Unlike Littlefinger, Claire is not an *unmotivated* frail dog. The problem, rather, is that when we keep our promises for reasons consistent with anticipatory guilt, we effectively only bind ourselves by binding ourselves to *someone else*. Again, that is

precisely what moral trust seems to require—that I give the vulnerability you entrusted to me, by accepting my promise on the basis of trust, priority over my personal desires.

This is the main reason that promises kept on the basis of moral trust prove to be weak or deficient on Nietzsche's analysis, and why moral trust is more tenuous than we often think. In general, I succeed in binding myself by promising precisely by binding myself *through you*. But promising is actually far more demanding than this. It requires of us that we be capable more fundamentally of *binding ourselves*, and this is far more difficult than we typically think. Consider the promises you've made to yourself throughout the years—perhaps to start a diet, to change a regrettable habit of character, or to begin a workout regimen—and how successful you were in binding yourself.[146] Perhaps that is because these promises are by comparison trivial compared to promises made on the basis of moral trust. Perhaps. Nietzsche's diagnosis, to the contrary, is that binding oneself requires that the will be self-determined or autonomous, and that autonomy is incredibly difficult to achieve. By contrast, promising on the basis of moral trust generally succeeds precisely by making the will heteronomous, insofar as we end up relying on the expectations of others to sustain the motivation to keep our promises.

[146] Jorah Dannenberg (2015) illuminatingly argues that promising to others depends on the ability to make and keep promises to oneself. The account of sovereign trust I offer below has been influenced by his work. However, I am now less confident in his argument. He claims, "By offering a promise, a person asks for the trust of another. In so doing, she in effect proclaims herself to be worthy of the trust she invites. For this proclamation to be sincere, she must also undertake to *be* the kind of person she proclaims herself to be … she must undertake to value, and to continue to value, her promise in a particular way" (2015: 173). I agree with what Dannenberg says here, but the "particular way" we keep our promises is relevant depending whether we promise on the basis of moral trust or sovereign trust. Moral trust requires only that I be moved favorably and adequately by your reliance, and while that does require that I "undertake to *be*" trustworthy, here that means only that I keep my promise out of a sense of moral obligation. It does not require that I keep my promise out of a motive of conscience or integrity—and *that*, I suggest below—is what Dannenberg's analysis so illuminatingly reveals.

4.2 Integrity and Autonomy

Mark Migotti (2013) and Bernard Reginster (2017) argue that sovereign trust is in an important sense *not* dependent on making promises to others. According to Migotti, "the philosophical focus of GM II: 1-2 is not the nature of promising in the narrow sense of making a pledge to do something for someone else, but the nature of pledging or committing oneself in general" (2013: 510). Similarly, Reginster claims "we might suppose that Nietzsche's chief preoccupation in discussing promising is the general idea of a voluntarily binding *commitment*, rather than the particular idea of a commitment made to someone else" (2017: 17). I agree with Migotti and Reginster that Nietzsche cared more fundamentally about the issue of binding oneself, the issue of "voluntary reliability" as Migotti characterizes it, as opposed to the ability to make and bind oneself to others, where this involves keeping one's promises out of a sense of moral obligation. I also agree, for this reason, Nietzsche cared more fundamentally about the ability to make promises to oneself.

Where I would *disagree* with Migotti and Reginster is that Nietzsche metaphor of "permitted promising" was not at all concerned with moral trust, as Migotti explicitly argues and as Reginster sometimes suggests. To be "permitted to promise," according to Migotti, a promiser must be a sovereign individual who keeps his promises out of a commitment to personal integrity (2013: 516-19). But if this is true, permitted promising cannot be a generally realized status that distinguishes human beings from other kinds of creatures, per GM II: 1. Migotti's interpretation thus falls prey to the second horn of the interpretive dilemma. Reginster's view, on the other hand, unsteadily straddles both sides of the interpretive dilemma I have brought attention to here. He maintains that sovereign individualism is a status that demarcates human beings from other creatures

(Reginster 2017: 7), *and* that sovereign individuals keep their promises out of a commitment to integrity, a "strength of will" that is "unbreakable" (Reginster 2017: 4). The last remark suggests that sovereign promising is a kind of ideal. However, I am inclined to conclude that Reginster's view falls prey to the first horn of the dilemma, since the general thrust of his argument is to show that sovereign individuals feel guilt, which allows them to identify non-prudentially with promise keeping. As he says, the sovereign individual "values being trustworthy—a man of his word" (Reginster 2017: 7).

In any case, I do not think that *only* sovereign individuals ought to be permitted to promise, or trusted. The reason for this is that I take the "memory of the will" to be a description of the *capacity* underwriting permitted promising, and as I argued in Section 3 most humans possess the ability to keep their promises in the absence of external incentives, in virtue of our ability to feel guilt. When we promise sincerely, we can be and generally are internally motivated to keep such promises, recognizing them to be moral obligations. It's just that when we keep promises in this way we do not promise "like a sovereign."

> This being who has become free, who is really *permitted* to promise, this lord of the *free* will, this sovereign—how could he not know what superiority he thus has over all that is not permitted to promise and vouch for itself, how much trust, how much fear, how much reverence he awakens—he "*earns*" all three—and how this mastery over himself also necessarily brings with it mastery over circumstances, over nature and all lesser-willed and more unreliable creatures? The "free" human being, the possessor of a long, unbreakable will, has in this possession his *standard of value* as well: looking from himself toward others, he honors or holds in contempt … he honors the ones like him, the strong and reliable … that is, everyone who promises like a sovereign, weightily, seldom, slowly, who is stingy with his trust, who *conveys a mark of distinction* when he trusts, who gives his word as something on which one can rely because he knows himself to be strong enough to uphold it against accidents, even "against fate"—: just as necessarily he will hold his kick in readiness for the frail dogs who promise although they are not permitted to do so, and his switch for the liar who breaks his word the moment it leaves his mouth. The proud knowledge of the extraordinary privilege of *responsibility*, the consciousness of this rare freedom,

> this power over oneself and fate, has sunk into his lowest depth and has become instinct, the dominant instinct:—what will he call it, this dominant instinct, assuming that he feels the need to have a word for it? But there is no doubt, this sovereign human being calls it his *conscience* ... (GM II: 2)

I am going to focus on three aspects of the sovereign individual Nietzsche mentions in the above passage, to show how sovereign promises differ from promises kept on the basis of moral trust. First, the sovereign individual is "free" or "autonomous," where this involves being "supermoral (for 'autonomous' and 'moral' are mutually exclusive)." Secondly, the sovereign individual is *trustworthy,* he "*earns*" trust when he promises, because he possesses a "long unbreakable will." Finally, he "knows the extraordinary privilege of *responsibility*," because responsibility has "sunk into his lowest depth and has become instinct," or because conscience is his "dominant instinct."[147]

Let's begin with the last idea. What does it mean to say that conscience is one's "dominant instinct?" Our investigation has revealed that conscience is a kind of "memory" that is uniquely responsive to remaining aware of and honoring one's obligations, be they non-moral or moral. When Nietzsche describes conscience in its mature form as a "memory of the will" (GM II: 1), as we have analyzed it in this chapter, it qualifies as an ability to extend practical commitment in the absence of external incentives. It is the "will's memory" precisely because it allows an agent to *bind himself* to a course of action, by reminding the agent of and generating the motivation necessary to, *sustain* the action into the future. Because it is characteristic of the memory of the will that this be achieved in the absence of external incentives, it presumes that the agent instead does so on the basis of values she accepts. As we have seen, *one* way we may do this is by keeping promises, obligations, or practical commitments out of a sense of

[147] I leave to the side Nietzsche's provocative remarks about "fate," but I take the significance of this to be that the sovereign individual does not use fate as an *excuse* when and if he fails to keep his promises, as we might imagine the ancient Greeks did.

moral obligation. In such cases we regard the action as a "personal obligation" that bears on our sense of worth, perhaps because we value being trustworthy, or honest, or because we value the fact that others rely on our goodwill. However, if Nietzsche thinks "autonomous" and "moral" are "mutually exclusive," I take it he does *not* think the sovereign individual keeps his promises for any of these reasons, all of which are ultimately rooted in guilt. So why does he keep them?

Well, as much as it would *seem* that Nietzsche is distancing himself from Kant by stating that morality and autonomy are "mutually exclusive," he would actually seem to be a lot closer to Kant than one might think. For it would seem that the sovereign individual keeps his promises out of something very much *like* a "motive of duty." When agents act from a motive of duty on Kant's analysis, in compliance with the categorical imperative, they act autonomously *and* morally because "reason alone" determines their action, because the action is not determined by the agent's "alien" subjective inclinations. Their action is autonomous because Kant identifies the will with reason, since "the will is nothing other than practical reason," a "capacity to choose *only that* which reason, independently of inclination, recognizes as practically necessary (Kant 1785/2012: 412). And so when "reason alone" determines action, the will determines *itself*, and is thus autonomous. Similarly, Nietzsche seems to be saying that the sovereign individual keeps his promises out of a motive of *conscience*, because conscience has "sunk into his lowest depth" and become his "dominant instinct." In that case the will's "memory" would be determining the will in analogous way, i.e., the will would be determining *itself*. However, in order for the will to be autonomous it must be "supermoral," because autonomy and morality are "mutually exclusive."

I do not take this to mean that one who acts autonomously cannot do what morality requires, or vice versa, that whenever one does what morality requires one necessarily acts heteronomously. The two are not "mutually exclusive" with respect to their requirements and outcomes. The point, I take it, is that the two are "mutually exclusive" with respect to the *motive* one must act from in order for one's action to qualify as "autonomous." This is where Nietzsche differs from Kant. He took the motive relevant to autonomy to be the motive of *duty*, a moral motive. Nietzsche takes it to be something else, a motive that is not essentially moral. In short, this means that, to be autonomous, one cannot honor values, obligations, or commitments from a sense of moral obligation, i.e., in a way that ultimately has its basis in guilt.[148] What, then, is this motive of conscience? And what would it mean if conscience alone determined the will?

We just observed that conscience is a faculty that "reminds" us of and helps us to maintain our values or practical commitments. If we imagine that I have a first-order commitment or value to, say, be trustworthy, conscience would be the faculty that *reminds* me that I value being trustworthy, and would *encourage* me to be trustworthy, whenever I am in danger of acting in ways that would reveal me to be untrustworthy. The typical response to violating such a first-order commitment or value on Nietzsche's view is guilt, the "feeling of personal obligation" (GM II: 8). Guilt, along with the faculty of bad conscience that produces it, are like the *gatekeepers* of our values in this way. If there is a form of conscience beyond this, a "highest form" (GM II: 3), naturally we might think that it manifests as a kind of *second-order* commitment to maintaining our first-order values and commitments. This is how Jorah Dannenberg describes the memory of

[148] I do not believe the sovereign individual is *incapable* of feeling guilt, only that the feeling of guilt would be quite *foreign* to him, for reasons I expand on below. Indeed, I believe the susceptibility to guilt is a condition of having a "memory of the will."

the will, as a "term of art for the sort of continuity or stability that a person undertakes to actively create and maintain within her system of values" (Dannenberg 2015: 164), or "an active effort to shape and maintain one's values" (Dannenberg 2015: 163). As I would characterize this, sovereign conscience manifests as a commitment to personal *integrity*, understanding integrity to be a state of wholeness or coherence within a person's values and practical commitments.

To see this, let's begin with a simplified model of integrity as coherence. According to John Bigelow and Robert Pargetter (2007), integrity is the capacity to exercise strength of will (2007: 42), and we exercise strength of will whenever we are successful in making our first-order desires conform to our higher-order desires. We need to supplement this account for Nietzsche's purpose because, as we have seen, willing on his analysis involves acting from an (internal) motive that has its source in the agent's values or practical commitments,[149] therefore giving conscience a privileged role in sustaining acts of will against sources of resistance. Therefore, strength of will is not simply the ability to make lower-order desires conform to higher-order desires; it is an ability to make lower-order desires conform to higher-order commitments, which stem from the agent's values. But what Bigelow and Pargetter conclude about integrity is very much consistent with a view we might attribute to Nietzsche:

> Integrity is a character trait. It comes in degrees. A person with integrity is one who can display strength of will not only when the temptations are slight but also when they are acute, not only on freak occasions but over a wide range of likely potential situations, and not only over short-term but also over long-term projects. (Bigelow and Pargetter 2007: 44)

[149] The same point is argued, albeit on slightly different grounds, by both Reginster (2017: 13) and Clark and Dudrick (2012: Chapter 7).

This is very close to Lanier Anderson's view of freedom or autonomy that I attributed to Nietzsche in Chapter 1. The "autonomous" person is well-ordered and unified, and thus able to overcome sources of resistance that would prevent the realization or maintaining of her goals, values, or practical commitments (her will).[150] And becoming autonomous on Anderson's analysis is itself a goal or a task, a "distinctive form of self-relation" achieved "through a kind of self-creation" (2013: 456). Autonomy also comes in degrees, according to him, because we can be more or less unified.

So, integrity is a *state* in which a person is whole or unified, specifically in the sense that her values are unified with one another and her actions consist with those values. Integrity is *realized* or achieved by exercising strength of will, by overcoming threats to our values and exercising will power. Therefore, the motive of conscience, as exemplified when it becomes an individual's "dominant instinct," is precisely a motive to *realize or maintain integrity*. As Dannenberg says, the memory of the will signals "a particular kind of strength of will required if one is to sustain one's voluntary undertaken commitments in the face of pressures that could otherwise lead to forgetting or abandoning them" (2017: 6, fn. 15). Because the memory of the will takes the form of a motive of conscience or integrity in the sovereign individual, and because he has achieved "autonomy" or "self-mastery," he can be trusted to keep his promises come what may. Unlike the rest of us, the sovereign individual "*earns*" the trust he invites, whenever he makes a promise. Because he, unlike the rest of us, possesses a "long unbreakable will."

---

[150] Consider the following remarks from Nietzsche's notebooks: "The multitude and disaggregation of impulses and the lack of any systematic order among them result in a 'weak will'; their coordination under a single predominant impulse results in a 'strong will': in the first case it is the oscillation and lack of gravity; in the latter, the precision and clarity of the direction" (WP 46).

4.3 Sovereign Trust and the Scope Problem

The sovereign individual keeps his promise from a motive of integrity or conscience, as opposed to keeping it because others are relying on his goodwill, or because he simply values being, say, trustworthy or honest. He *does* value being trustworthy or honest, and the fact that others are relying on his goodwill *would* be a consideration that moves him favorably. However, these are *not* the reasons he keeps his promises, because he does not keep his promises out of a sense of moral obligation.[151] On Nietzsche's analysis most of us do *not* act from this motive of integrity or conscience. If we keep our promises for reasons that our consistent with moral trust at all, we general keep them on the basis of these first-order values or commitments, fidelity to which is paradigmatically secured by my binding myself to *another* and keeping my promise out of a sense of moral obligation. This, I have argued, is the basis of *moral trust*; while it is an achievement relative to other *animals*, it is *not* an achievement relative to what *humans* can aspire to and attain. The reason for this is that, whenever we keep promises, or honor values or commitments out of a sense of moral obligation, we *fall short* of the greatest task bequeathed to human beings—the task of "becoming a self" (HA II: 366) or becoming autonomous. We therefore do not realize the "consciousness of power and freedom" that humans can aspire to, the "feeling of the completion of man himself" (GM II: 2).

To expect a sovereign individual to keep his promises on the basis of moral trust would amount to a kind of affront or offense, a failure to extend the higher degree of regard he is due when he makes promises. Many scholars stress this by interpreting

[151] Is integrity not relevant to morality, one might wonder? Note, we would not seem to have a general moral obligation to realize integrity or autonomy, or to act from a motive of integrity. Integrity is most relevant to morality perhaps and only to the extent that when one *blames* another, it would be good for her to do so from a non-hypocritical standpoint. Integrity is arguably necessary to the end of avoiding hypocrisy, and so for blame, but it is not a more general moral requirement.

dürfen as "entitlement" or "right" to make promises, instead of merely having the "permission" to do so. Dannenberg does so explicitly to capture "the sort of privilege, power, and authority that accrue to a person of her word" (2017: 1, fn. 1). No Nietzsche scholar I am aware has considered the implications of interpreting *dürfen* in this way, though Dannenberg has. According to him, "rightful promisers possess something more along the lines of what many would designate a 'claim right,' corresponding to an obligation on the part of others to take them at their word when they see fit to give it" (Dannenberg 2017: 1, fn.1). Consider me skeptical, at least as a reading of Nietzsche. (To be clear, Dannenberg is not trying to give a faithful rendering of Nietzsche.)

If we have an *obligation* to trust sovereign individuals whenever they give their word, sovereign promises not only start to look a lot like threats, but, more importantly, honoring them by accepting their word would seem to place us and him squarely back within the realm of guilt and moral responsibility. The sovereign individual is supposed to be beyond this. I have nonetheless tried to argue that sovereign trust has its basis in a kind of entitlement or regard. I would put it like this: sovereign individuals alone are *trustworthy* because they and only they prove capable of *binding themselves* when they promise. They "*earn*" the trust they invite by making promises, and so we owe it to them to give them that regard, if we should choose to accept their promise. We do this by taking them at their word, even when keeping one's word is onerous and borders on impossible. Whenever *he* trusts (and he does so sparingly), he "conveys a mark of distinction" (GM II: 2). Accordingly, he would expect or demand that distinction in turn.

I take it the above considerations do not imply that the sovereign individual is immune to guilt, the reactive attitudes, and moral responsibility, though it does mean that these are quite *foreign* to him. Precisely, he is "beyond" or "above" the moral

reactive attitudes in that he does not need the prospect of guilt, the expectations of others, along with the second and third-personal reactive attitudes, to honor his promises, values, or commitments. He does not need these things most of us *do* need because he values something more fundamental, and which takes these under its purview. He values being *coherent* or *whole*, a person of integrity, "free" and "autonomous." Making a promise is an invitation to fracture that coherence or wholeness, because breaking the promise would reveal himself to be what he finds most contemptible, namely, *weak-willed* and *disunified*. So we can now see why the idea of "responsibility" he exemplifies is both an "extraordinary privilege" and exceedingly "rare" (GM II: 2). It is so because he keeps his promises, and remains true to his values and practical commitments more generally, out of a motive of conscience or integrity. He does not merely *possess* a "memory of the will" and keep his promises out of a sense of moral obligation, he represents the *exercise* of that capacity to the highest degree. He is an *ideal* of responsibility because conscience is his "dominant instinct."

Two lessons follow from the model of trust the sovereign individual provides. First, if we are to "promise like a sovereign," we must do so "weightily, seldom, [and] slowly" (GM II: 2). The reason for this is that making a promise is the act of binding oneself, and this is no minor or easy feat considered apart from moral obligation and guilt. Applied to the case of Claire, this means that integrity would demand of her either that she not agree to attend the birthday party, or, as I think more plausible, that she not promise to meet her friends after having promised to attend the party. One or both of these promises was made flippantly from the perspective of sovereign trust, and she was therefore *unworthy* of being a person who is "permitted to promise" in the stronger sense, though this did not impugn moral trust, her being "permitted to promise" or

relied upon from the participant stance. Secondly, sovereign trust is compatible with moral trust in one sense, because sovereign individuals will not let us down if we rely upon them from the participant stance. But in another and arguably more important sense, sovereign trust is *not* compatible with moral trust. For if one keeps a promise out of a motive of integrity, then she will not experience guilt, and if she *does* experience guilt, then she was not keeping her promise out of a motive of integrity. Because in that case she did not endeavor to *bind herself* when she made the promise.

V. Conclusion

When Nietzsche first introduces the idea of "conscience" that plays a central and shifting role throughout GM II, he describes it as the "memory of the will," as a faculty that enables us to bind ourselves, or to extend practical commitment in the absence of external incentives. He then describes its instantiation in the sovereign individual, where conscience is exemplified in its "highest form" (GM II: 3) as a commitment to personal integrity. Nietzsche then goes all the way back to the beginning, to humanity's "prehistory," where he explains the conscience originated as a memory of customary rule-prohibitions, or "I will nots" (GM II: 3). Many aphorisms later he then explains the origin of the most eventful and problematic form of conscience—"bad conscience" (GM II: 16). He leaves it up to the reader to connect all these disparate strands and pieces connecting his "long history and metamorphosis" (GM II: 3) of conscience. Our investigation has yielded the following, I think plausible, summary.

Humans are superior to other creatures, indeed "responsible," because they *possess* a "memory of the will." This makes them "permitted to promise" in the weaker sense that they can be relied upon from the participant stance, because in general they keep their sincere promises out of a sense of moral obligation. This is ultimately rooted

in their being morally responsible agents who are capable of feeling guilt, which depends on the development and exercise of bad conscience, the ability to feel a "pang of conscience" (GM II: 14). Bad conscience is not just a passive memory of "I will nots" (GM II: 3), because this latter form of obligation depends on being held to an expectation of conformity by somebody, somewhere, who has power to punish. The development of bad conscience made humans *responsible*, and not merely reliable, because it allowed us to *hold ourselves* to expectations, and those expectations became moral or personal obligations once they became attached to judgements of guilt. Finally, the sovereign individual represents an *ideal* of responsibility, connected to Nietzsche's conception of autonomy, because he does not keep his promises, or keep his values or commitments, out of a sense of moral obligation. In him conscience has become its own motive, his "dominant instinct," and the significance of this is that he is capable of *binding himself* when he promises. He is "permitted to promise" in the stronger sense that he is actually *trustworthy*; he is "permitted to vouch for *oneself*, and with pride" (GM II: 3, emphasis mine).

References

Works by Nietzsche

A: *The Antichrist*. 1888. In W. Kaufmann, trans. and ed., *The Portable Nietzsche*. New York: Viking Press, 1954

BGE: *Beyond Good and Evil*. 1886. In W. Kaufmann, trans. and ed., *The Basic Writings of Nietzsche*. New York: Modern Library Edition, 2000.

GM: *On the Genealogy of Morality*. 1887. Clark, Maudemarie, and Swenson, Alan J, trans. Indianapolis: Hackett Publishing Company, 1998.

GS: *The Gay Science*. 1882/1887. W. Kaufman, trans. New York: Vintage Books, 1974.

HA: *Human, All Too Human*. 1878. R.J. Hollingdale, trans. Cambridge: Cambridge University Press, 1996.

TI: *Twilight of the Idols*. 1888. In W. Kaufmann, trans. and ed., *The Portable Nietzsche*. New York: Viking Press, 1954

WP: *The Will to Power*. W. Kaufmann trans. New York: Random House, 1968.

Other Works

Acampora, Christa. 2006. "On Sovereignty and Overhumanity: Why it Matters How We Read Nietzsche's Genealogy II, 2," *International Studies in Philosophy* 36: 127-45.

Anderson, R. Lanier. 2013. *Nietzsche on Autonomy*. In Gemes and Richardson, ed., *The Oxford Handbook on Nietzsche*. Oxford: Oxford University Press.

Baier, Annette. 1986. "Trust and Antitrust." *Ethics* 96 (2): 231-60.

Bigelow, John and Pargetter, Robert. 2007. "Integrity and Autonomy." *American Philosophical Quarterly* 44 (1): 39-49.

Clark, Maudemarie and Dudrick, David. 2012. *The Soul of Nietzsche's Beyond Good and Evil*. Cambridge: Cambridge University Press.

Dannenberg, Jorah. 2017. "Promising by Right." *Philosopher's Imprint* 17(22): 1-18.

__.2015. "Promising Ourselves, Promising Others." *Journal of Ethics* 19: 159-183.

Hieronymi, Pamela. 2008. "The reasons of trust." *Australasian Journal of Philosophy*, 86:2, 213-236.

Hatab, Lawrence J. 2009. "Breaking the Contract Theory: the Individual and the Law in Nietzsche's Genealogy." In Siemens, H.W., and Roodt, V., eds., *Nietzsche, Power, and Politics: Rethinking Nietzsche's Legacy for Political Thought*. Berlin: W. de Gruyter, 169-88.

Holton, Richard. 1994. "Deciding to Trust, Coming to Believe." *Australasian Journal of Philosophy,* 72: 63-76.

Jones, Karen. 1996. "Trust as an Affective Attitude." *Ethics* 107: 4-25.

Kant, Immanuel. 1785/2012. *Groundwork of the Metaphysics of Morals*. M. Gregor and J. Timmerman, trans. Cambridge: Cambridge University Press.

Leiter, Brian. 2015. *Nietzsche on Morality* (3rd Edition). London: Routledge.

__.2011. "Who is the 'Sovereign Individual'? Nietzsche on Freedom." In Simon May ed.,*"Nietzsche's On the Genealogy of Morality": A Critical Guide*. Cambridge: Cambridge University Press.

McGeer, Victoria. 2008. "Trust, Hope, and Empowerment." *Australasian Journal of Philosophy* 86 (2): 237-254.

Migotti, Mark. 2013. "A Promise Made is a Debt Unpaid: Nietzsche on the Morality of Commitment and the Commitments of Morality." In Gemes and Richardson, ed., *The Oxford Handbook on Nietzsche*. Oxford: Oxford University Press.

Riccardi, Matthias. 2015. "Inner Opacity: Nietzsche on Introspection and Agency." *Inquiry* 58 (3): 221-243.

Reginster, Bernard. 2017. “What is the Structure of Genealogy of Morality II?” *Inquiry* 61 (1), 1-20.
___. 2011. “The Genealogy of Guilt.” In May, Simon ed., *Nietzsche’s “On the Genealogy of Morality”: A Critical Guide*. Cambridge: Cambridge University Press, 56-77.

Strawson, P.F. 1962. “Freedom and Resentment.” In G. Watson ed., *Free Will: Second Edition*. Oxford: Oxford University Press, pp. 72-93.

Walker, Margaret Urban. 2006. *Moral Repair: Reconstructing Moral Relations after Wrongdoing*. Cambridge: Cambridge University Press.