Nietzschean Autonomy and the Meaning of the "Sovereign Individual"[1]

R. Lanier Anderson

Department of Philosophy, Stanford University

This paper has two goals—a narrow one which I take myself to achieve, and a more ambitious one toward which I can make only preliminary suggestions. The limited goal is to resolve an interpretive dispute over Nietzsche's intentions in the *Genealogy*'s description of the "sovereign individual," a character type whose features turn out to have important bearing on Nietzsche's distinctive conceptions of conscience, promising, and what it is to take responsibility for oneself. The more ambitious goal is to characterize what Nietzsche means by autonomy and to assess what we might have to learn from his conception. The paper's basic idea is that the meaning of the sovereign individual emerges clearly in the light of a distinction from Bernard Williams between two senses of responsibility—one sense tied to voluntary action, the other to an ambitious conception of responsible agency. What we learn about responsibility from Williams then illuminates how we should understand Nietzschean autonomy, and what philosophical purposes his conception can (and cannot) serve.

1. The Textual Puzzle

Nietzsche's description of the sovereign individual has generated a surprising amount of controversy in the recent literature. It seems obvious (I, for one, took it to be obvious from the first time I read the *Genealogy*) that Nietzsche introduces this character in order to praise it by

contrast to a life dominated by the moralized notion of guilt criticized in the ensuing Second Treatise. This first-pass reaction is still shared by most readers,[2] and it finds its most salient source of support in Nietzsche's evaluative language—as so often with him, high-flown rhetoric is hardly in short supply. Here is the central text:[3]

> If… we place ourselves at the end of the enormous process, where the tree finally produces its fruit, 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 its tree the *sovereign individual*, like only unto himself, liberated again from the morality of custom [*Sittlichkeit des Sitte*], autonomous and super-moral [*übersittliche*] (for 'autonomous' and 'moral' [*sittlich*] are mutually exclusive), in short, the human being with his own independent long will, the human being who is *permitted to promise* [*der versprechen darf*]—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 [*Vollendungs-Gefühl des Menschen überhaupt*]. [*GM* II, 2]

The broad rhetorical tone seems obvious: the sovereign individual is, well, *sovereign*—and what is more, the "ripest fruit" of a long process, the privileged end to which the entire formation of society is to be seen as a merely instrumental means, as well as "proud," unique, independent and free, powerful and autonomous, even (Nietzsche doesn't hold back) the completion (or consummation, perfection) of humanity itself! The follow-on passage reports that this type of person "knows" his "superiority" over all less developed others, that he "'*deserves*'" the trust, fear, and reverence he inspires, that he is "strong and reliable" enough to attain "mastery over

himself"—and even over "circumstances" and "fate"—and that on the basis of all this, he has the "extraordinary privilege of *responsibility*," which justifies his "permission to promise" (*GM* II, 2). The rhetoric may be overblown, it may be off-putting in its anti-egalitarianism and "muscular" taste, but in those very respects, it seems entirely in the spirit of Nietzsche's usual patterns of praise. What could be clearer?

Dissent from the obvious, first pass reading of Nietzsche's evaluative stance has arisen from two quarters. The initial doubts arose from scholars—Lawrence Hatab (1995; 2008; 2009), followed by Christa Acampora (2004; 2006; 2013, 132-9)—pursuing a Nietzschean political theory based on his agonistic conception of social relations.[4] They observe that Nietzsche introduces the sovereign individual not as a successor of altruistic morality in general, but instead as someone who transcends the more primitive "morality of custom [*Sittlichkeit der Sitte*]" (*GM* II, 2; compare *D* 9), which roots normative demands in the unquestioned authority of *custom* rather than rational justification or altruistic sentiment. Hegel, of course, had marked a strong distinction between *Sittlichkeit* and the abstract, rational demands of *Moralität*, criticizing Kant for leaving no room for the former. If we assume that similar weight can be rested on an implicit distinction between *Sittlichkeit* and *Moralität* in Nietzsche, it can suddenly seem that the sovereign individual is not a "super-moral" character who escapes the problematic features of morality in general, but instead exactly the rational, autonomous subject of standard modern moral and political theory—independent simply from tradition and custom, not from morality itself. After all, he is distinctive primarily in his capacity to honor promises, including, Hatab (2008, 80) suggests, the founding promise by which he enters the social contract.[5] This makes

him seem like the free and responsible agent needed for (and subject to) the demands of morality. Accordingly, the Nietzschean should be skeptical, not enamored, of the sovereign individual.

There is also an underlying philosophical motivation to promote such doubts. Hatab and Acampora make much of a skeptical strand of Nietzschean texts about the self, which depict it as a disunified, irreducibly plural collection of competing drives.[6] This strand fits with the agonistic conception of Nietzschean politics; in the spirit of the old Platonic analogy, the multiple, internally conflicted self parallels agonistic competition within the state. But the sovereign individual is highly unified through his own self-mastery, so if Nietzsche genuinely meant to idealize this type, then his apparent rejections of unified selfhood would need to come in for some qualification. By raising doubts about the laudatory tone highlighted in the standard reading, Hatab and Acampora can avoid the need for such qualifications.

A second source of recent skepticism is Brian Leiter (2011), whose relentlessly deflationary naturalist interpretation of Nietzsche places heavy emphasis on fatalistic denials of freedom of the will. In that spirit, Leiter opposes interpretations that rely on the standard reading of *GM* II, 2 to highlight the sovereign individual's *autonomy*—(scholars often appeal to this passage to identify a positive, compatibilism-friendly conception of freedom as self-control, alongside Nietzsche's acknowledged rejection of metaphysical free will).[7] Leiter arrives (apparently independently) at some of the same textual observations as Hatab and Acampora, including 1) the point that the sovereign individual transcends morality (only) in the sense of custom (*Sittlichkeit*), 2) the thoroughly naturalistic character of Nietzsche's surrounding discussion, which traces the capacity to promise to a deep history of memory training (noted by

Acampora 2006, 148-50, 157), and 3) the alleged similarity between the sovereign individual's "*permission to promise*" and the modern individual's realization of virtues apt for commercial society. He also shares the suspicion that readers who hear Nietzsche's description as straightforward praise are being seduced into a moralized misreading that sanitizes Nietzsche's elitism and hard naturalism for the benefit of our soft, liberal sensibilities (Leiter 2011, 119).

For Leiter, however, more of the argumentative weight rests on the claim—based on *other* frankly necessitarian texts—that Nietzsche never could have intended non-ironic praise of autonomy. Given Nietzsche's uncompromising fatalism, any "sovereign individual" who took his self-mastery to reflect his own control over his life, as opposed to a fated condition rooted in his drives, must be "delusional" (Leiter 2011, 109). In this mood, Leiter is tempted to turn the very rhetorical intensity of Nietzsche's depiction against the idea that the praise is seriously intended, detecting "ridiculous hyperbole" (Leiter 2011, 103) in the connection of such a high-flown encomium to the (pathetic) bourgeois virtue of honoring contractual obligations (Leiter 2011, 108). Still, he concedes that there is probably more to the passage than ironizing the bourgeoisie—mainly (now *contra* Hatab and Acampora) in its valorization of self-mastery and coherence of character. Nevertheless, he insists that such virtues have no real connection to freedom, as *philosophically* understood.[8] On the contrary, they are fated behavioral dispositions, and Nietzsche connects them to "autonomy" in this context as part of a rhetorical program of "persuasive definition" designed to exploit the positive associations of 'freedom' and 'autonomy' toward ends wholly incompatible with serious philosophical understandings of those concepts (Leiter 2011, 103, 112, 117-18).[9]

These arguments are unlikely to persuade the scholars Leiter is criticizing. The whole point of efforts to locate a compatibilism-friendly conception of autonomy in Nietzsche was that such a notion would *not* be incompatible with his necessitarianism. Thus, texts expressing fatalism will exert *no* pressure on them to follow Leiter's dismissal of Nietzsche's frequent praise of freedom (in some sense) as so much "persuasive definition."[10] That said, one could adduce on Leiter's side the philosophical observation that it is hard to see the attraction of the compatibilist program itself for someone with Nietzsche's presuppositions. Compatibilism is motivated by the felt need to reconcile real and binding moral responsibility with a certain (determinism-involving) conventional wisdom about our best scientific picture of the world. But Nietzsche is not at all concerned to save moral responsibility—on the contrary, he repeatedly denounces the whole institution as a dishonest power-play by a priestly class (e.g., at *TI* VI, 7; *GM* I, 13)—and he often seems entirely willing to reject standard, natural law-based determinism, as well (*GS* 127; *BGE* 21, 22). So, if the standard compatibilist program holds no attraction, why should we think Nietzsche adopts some super-subtle conception of freedom appropriate for such a view? In the end, we appear to be at an impasse.

In my view, it is no accident that controversy over the meaning of the sovereign individual cannot be settled by closer reading of the text alone. Given the peculiarities of Nietzsche's authorial identity—his writerly ambition, his subtlety, his frequent indirection and proclivity for strategies of audience-partition (esoterism)—the passage's evaluative language is bound to be open to multiple interpretations, as Leiter's ironic reading reminds us. Meanwhile, the other terms most relevant to the dispute—'freedom' and 'autonomy' themselves, of course,

but also 'individual,' 'self-mastery,' 'morality,' 'responsibility,' and the 'right' (or 'permission') to promise—capture highly abstract, philosophically-loaded notions that get much of their detailed content from the surrounding theory within which they operate. They permit too many degrees of freedom in interpretation to settle the dispute by their mere appearance on the scene, and Nietzsche's use of them in the context is not designed to cut down that flexibility by careful technical specification of the intended conceptions, but instead to advance other agendas. If we are to understand the sovereign individual, we will have to probe Nietzsche's underlying philosophical motivations. My basic suggestion is that improved optical resolution on the pertinent terrain emerges from a distinction in one of Bernard Williams' incisive, if somewhat telegraphic, papers from the era of *Shame and Necessity*.

2. Voluntary action and responsible agency

That paper is "Voluntary Acts and Responsible Agents" (Williams 1995 [1989]), and the relevant distinction separates what makes an action voluntary (such that we can hold the agent responsible), from what enables an agent to take responsibility for herself—not just for this or that action, but also for her character, for her role within her social groups, for her life overall. In a slogan, the distinction is one between the basis for *holding* someone responsible and the basis for someone's *taking* responsibility. In typical Williams style, the point of the paper is to suggest that this quick distinction has more to it than meets the eye.

Not all philosophers would accept that the difference rests on a *deep distinction* at all. Indeed, Williams frames his paper as a resistance action against a philosophical program aiming

to explain the the role of voluntariness in holding people responsible by connecting it to a more basic notion of responsible agency (taking responsibility).

A quick and dirty version of that strategy holds apparent promise, as we can see with a bit of setup. Suppose we define voluntary acts as those under an agent's control, in the sense that they are intentional and connected in the normal way to whatever deliberation she undertook—e.g., free from defects like false beliefs about key relevant facts, or "deviant causal chains," or the operation of a compulsion, or akrasia that decoupled deliberation from action, or the like.[11] Still, one could worry that the voluntariness of an act was neither necessary nor even sufficient for holding the agent responsible. Against the *necessity* claim, recall that we are willing to deploy the Strawsonian reactive attitudes and hold people responsible even for acts that are *unintentional*. In many standard cases of negligence, people who did not know what they were doing are nevertheless answerable on the grounds that they *should* have known, but there are also more interesting cases where, even though the agent *could not* realistically have known the effects of her activity, she should have known better than to do any such thing—Williams' (1995 [1989], 25-6) example is the act of a dignitary recruited to cut the ribbon opening a nuclear power plant, who fiddles with some knobs in the control room and causes a radiation release. The claim that voluntariness is *sufficient* to place an act in the right category to hold the agent responsible carries greater initial plausibility, but it cannot be right either. Children (and arguably non-human animals) can act voluntarily, but at least in many cases, they cannot be held responsible—if we replace the dignitary in the nuclear plant example with a five-year old, she cannot rightly be held accountable, though someone *else* (some adult) may be responsible for a

failure of adequate supervision. But now the quick and dirty strategy can be wheeled in to salvage what was plausible about the sufficiency claim from these counterexamples: voluntariness could still be enough to hold an agent responsible, as long as it's the *right kind of agent*—a *responsible* agent. So on this strategy, the connection between voluntariness and holding responsible gets explained *via* the underlying notion of responsible agency.

This version of the idea *is* "quick and dirty"; it seems disappointingly uninformative since it assumes a notion of responsibility in specifying "the right kind" of agent as "responsible," and the account also faces a more interesting challenge about how to *focalize* the agent's responsibility onto the particular act. As Williams (1995 [1989], 26) pointed out, the notion of an "act for which the agent is responsible" can't simply be analyzed as "voluntary act of a responsible agent," because we could still doubt whether everything the person does must fully express her responsible agency, and thus, whether her *general* responsibility "translates down" to the particular action that concerns us. These limitations can be addressed, however, by appealing to a more ambitious conception of responsible agency, which might explain what that responsibility consists in and imply answers about when it is in effect. Unsurprisingly, that ambitious idea turns out to be the real focus of Williams' doubts.

For this particular article, Williams takes his point of departure from a justly influential paper by Terry Irwin (1980), who aimed to trace voluntary action back to responsible agency as a way of saving Aristotle's account of responsibility from a traditional objection based on the same observation about voluntary action by (non-responsible) children and animals. But the basic philosophical strategy has wider attraction and has been deployed in a variety of modern efforts

to root the notion of freedom that matters for responsibility in deeper structures of individual character or agency.[12] The distinctive idea in these otherwise quite various treatments is the notion of responsible agency at their core, which is what Williams (1995 [1989], 27) calls an "*ideal-bearing* account"; that is, it relies on an *aspirational* conception of taking responsibility for yourself, which can be realized to a greater or lesser degree, and which, while it may be *minimally* acknowledged (in the breach, at least) by any normal, well socialized adult, will be fairly demanding in anything close to genuine realization, projecting a relatively high (perhaps even unattainable-seeming) ideal as the proper goal of our efforts at self-control. The demanding thought arises from the insistence that taking responsibility for yourself must involve not only deliberation about what to do and how to order your desires in case of conflict, but also deliberation about which desires you ought to have in the first place, so that your desires themselves and even your character are responsive to practical reason. Irwin's version starts from the modest-sounding idea that a responsible agent needs the capacity to deliberate and must be open to reasons in relevant circumstances, but he is quickly driven (by reflection that such deliberation should not be limited to *mere means*) to the more ambitious thought that a truly self-controlling, responsible agent should not be the *victim* of her desires but must be capable of "doing something about them," where that turns out to involve rationally coordinating any given desire with "all his aims" together, and thus, forming a deliberative conception of the final aim of happiness (Irwin 1980, 128-31).[13] In Irwin's version, the ideal-bearing account is fairly rationalistic, but exclusive reliance on practical rationality is not a necessary feature. The general strategy can be just as ambitious and ideal-involving for other philosophers, who orient

the basic structures of character that are the target of self-control around self-governing policies, or values, or wholeheartedness, or stable satisfaction, or full self-intelligibility, or a necessary practical identity,[14] rather than around a reason-based power of practical wisdom. The goal is still self-regulation of the basic structures of character or the fundamental patterns of desire that contribute to it, and the result will be an ideal-bearing conception that aspires to a more or less ambitious aim of ordering our life overall.

To avoid begging any questions about responsibility, Williams characterizes the ideal as one of "*mature agency*" (Williams 1995 [1989], 28). This formulation has the dual advantages 1) of not immediately assuming a privileged place for rational deliberation in the picture (since it might be part of maturity to ensure that deliberation remains in its proper role, and that might be limited), and 2) of connecting the idea more firmly to its local dialectical purpose—viz., specifying what makes normal adults responsible in a way that children or animals are not, despite whatever capacities for voluntary action they exercise. As we'll see, the characterization also sets up a powerful aphoristic formulation of one side of Williams' argument against the philosophical program he is exploring.

But before moving to that argument, it is useful to explore the potential connection between ambitious mature agency and responsibility for particular voluntary actions. The connecting idea is a conception of freedom understood in terms of control. Being responsible for a particular act is tied to that conception because it is plausible to think someone responsible for an action if doing it was in her control, and what drove the ideal of mature agency was a similar idea of control turned inward toward the self—a form of *self*-control that enables the agent to

order and prioritize her practical attitudes and manage them in the face of obstacles. We hold people responsible as part of a social expectation that people should control themselves when they act, and such control looks like the minimal version of a deeper form of reflective self-control that is built into the ideal of mature agency when it emphasizes the coordination of drives, desires, and values across one's practical life as a whole. Thus, the attraction of making the ideal of mature agency fundamental to ordinary responsibility arises from the sense that the ideal is the further development of a capacity needed by anyone who can deliberate—and who will therefore need to assess the place of one desire *vis à vis* others within her psychic economy. Such basic deliberative rationality is supposed to be key to the form of self-control that makes me responsible, and so my being held responsible is rooted in my approximation—more or less, and the more the better—to the ideal of mature agency (see Williams 1995 [1989], 28-9).

Williams doubts that this attractive connection can withstand scrutiny, and his argument, while telegraphic and suggestive, strikes me as powerful. It is related to his general interest in the Edward Craig-inspired strategy of allowing our interpretation (and perhaps *reconfiguration*) of key philosophical ideas to be guided by a careful analysis of what genuine functions they perform and what social and psychological needs they fulfill.[15] Williams uses such an analysis to identify a basic mismatch between being responsible for an act and the ideal of mature agency—a mismatch running in both directions that undermines the apparently plausible supposition that they coalesce through connection to the same notion of freedom. Holding others responsible is a practice that receives (and must receive) its rationale and central domain of application from our efforts to regulate social relations among people who do not know or have any special

importance for each other. Its home territory is in the theory of justice and political relations, concerning social coordination among anonymous members of a public. Thus, what we aim to get hold of when we hold someone responsible is not who she is to her nearest and dearest, but a thinner layer of her agency—rooted just in a general awareness of social expectations and a broad disposition to conform, a level of foresight and self-control that allows people "to avoid unnecessary collisions with the law or with each other" (Williams 1995 [1989], 29). Since the point of responsibility in this sense is social coordination, not individual self-perfection or intimate "character friendship," *holding* someone responsible seems to have no need to invoke the agent's substantive conception of happiness, or her approximation, however distant, to the ideal of a maximally coherent and self-endorsed character for which she *takes* responsibility.

Williams deploys his mismatch argument to add flesh to the bones of this suspicion. It is not only that reflection about the actual socio-psychological roles of holding responsible and of the ideal of mature taking responsibility place them further apart than they seemed. On reflection, the two notions of responsibility stake their claims in different domains altogether. Holding responsible is about social coordination and control. Taking responsibility is about the formation of the person—within, but also as defined *against*, her social roles and contexts.

Thus, on one side, the development of mature agency and self-understanding simply lacks the kind of *standing* within the public domain (where we hold one another responsible) that our ordinary intentions and plans have. Particular ordinary plans can intersect and interfere with those of others, and for just that reason they are of concern to the practice of holding one another responsible. But for those purposes, it does not matter how, or even *whether*, each person's

various plans coalesce within an overall life-plan. As Williams nicely puts it, "the maturity or self-understanding of the well-formed agent never attains a necessary claim on the attention of the public, in the way the person's intentions may have on it in a system of mutual acknowledgement" (Williams 1995 [1989], 32). Our mature self-understanding is not what we are held responsible for, and what we *are* held responsible for may or may not contribute importantly to it.

Even more telling is the converse mismatch on the other side. Any agent who has gone at all far down the path of maturity toward a coherent character will see herself in, and feel compelled to *take* responsibility for, not just her voluntary acts, but a much wider swath of her life, including unforeseen and unintended aspects of what she does, and even some substantially unchosen aspects of her character and social situation with which she identifies. Her character and her life, Williams rightly insists, go beyond her deliberative choices, and what she wants to take responsibility for is *who she is*, not just this or that action. Williams alludes to his discussion in *Shame and Necessity* (Williams 1993, 66-74) of examples like Oedipus, or like Ajax, who perpetrates his ridiculous attack on the flock of sheep in a divinely induced state of insanity, but who would never even *form*, much less consider or accept, the thought that his not knowing what he was doing was the kind of excuse that could restore his honor or remove his shame. What matters here is not whether he acted voluntarily, but who he is: no one who expects of the world what Ajax expects—the recognition and honor accorded a genuine hero—could go on expecting it in the face of such humiliation, and he is unwilling to live a life without that expectation. Williams' basic point, however, goes beyond such dramatic cases, and also beyond

cases of mere negligence, precisely because mature agency is such an ambitious ideal. It aims at an order covering the person's whole life (or much of it), and connecting her practical activity to a conception of that life. The form of taking responsibility associated with the ideal must therefore stretch out far beyond the domain of voluntary action. As Williams beautifully puts it, in the aphoristic formulation I foreshadowed above, "No conception of public responsibility can match exactly an ideal of maturity because… to hold oneself responsible only when the public could rightfully hold one responsible is not a sign of maturity" (Williams 1995 [1989], 32).[16]

So mature taking responsibility extends far beyond the voluntary and that for which we can hold each other responsible in one direction, while in another, it is very doubtful that every action for which we could rightly be held responsible by someone will have sufficient importance in our life overall to matter for the ideal of mature agency. Moreover, as we saw, the particularities of the individual character I fashion for myself need not gain public importance simply from the weight I attach to them, so they fail to engage the system of social control and coordination in which holding responsible gains its purchase and its point. Williams is right: the two kinds of responsibility—holding responsible and taking responsibility—are different ideas, with different ethical functions and different interactions with other ethically important ideas, including freedom and self-control. Neither explains the other in a deep way.

3. The Sovereign Individual Revisited

My next suggestion is straightforward, and you will have anticipated it. The sovereign individual was special because of his "extraordinary privilege of *responsibility*" (*GM* II, 2). I

propose that the scholars disagreeing about what this means are implicitly relying on different types of responsibility. With Williams' distinction in hand, we can sort out these different senses. Once we do, it turns out that the standard, first pass reading of the passage, which accepts the evaluative language in its *non*-ironic force, clearly makes for a better interpretation.

Williams' distinction is a natural fit for Nietzsche. Following Nehamas (1985), many scholars have recognized Nietzsche's interest in self-fashioning or self-creation, and the idea of taking responsibility for yourself as a mature agent has clear relevance for any such project. More important, insisting on a *distinction* between *taking* responsibility and the social practice of *holding* responsible opens space for *criticisms* of holding responsible that are central to Nietzsche's agenda in general, and to the *Genealogy*'s Second Treatise in particular. Given the distinction, the self for which one takes responsibility can remain as a locus of value while the separate practice of holding responsible comes under attack. The content of those attacks can appeal to the damage done to that self by morality's practices of holding responsible, and Nietzsche's genealogy of guilt seems designed to follow just such a strategy.[17]

So in reading the *Genealogy* we <u>should</u> distinguish, with Williams, between taking responsibility and holding responsible. The question then becomes, which idea is under consideration when Nietzsche opens section 2 by declaring, "Precisely this is the long history of the origins of *responsibility*," and proceeds to introduce the sovereign individual's "extraordinary privilege of *responsibility*" (*GM* II, 2)? Is this character defined by being held (or holding) responsible in relations with others, or by taking responsibility for himself? Viewed from this angle, the text is clear. The sovereign individual's main distinguishing characteristic is that he is

no longer bound by the customary moral demands of the society from which he emerged: he is "free again from the morality of custom" and is "autonomous and super-moral" in just that sense. Nor is there any hint in this section of a new social or moral order (post-morality-of-custom) that holds these individuals responsible; on the contrary, all the focus is on the individual's "mastery over himself" and the way *he* looks out "from himself toward the others" in order to "honor or despise." Nietzsche does highlight the individual's "*conscience*"—that is the name he gives to the "proud knowledge" of his responsibility now made instinctual in him: "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*…" But Nietzsche introduces no others onto the scene who would call upon that conscience to hold the sovereign individual to account. Indeed, their absence is no accident, since others' *holding* him responsible would compromise his defining sovereignty.[18] In the end, the sovereign individual's conscience is all about *holding himself* responsible—i.e., something like *taking* responsibility, in the sense of Williams-style mature agency—not about a social practice of holding responsible (quotations from *GM* II, 2).*B

The skeptical readings, meanwhile, clearly must assume the opposite—that the sovereign individual's responsibility is to be taken in the sense of *holding* people responsible within impersonal social relations. That was the light in which this "history of the origins of *responsibility*" could seem to Hatab to place the individual and his promising within the traditional social contract of modern political theory; it is why Leiter insists that the sovereign individual's self-control would, unless it is being ironized or made fodder for persuasive definition, involve Nietzsche in commitment to a type of freedom that would underwrite *moral*

responsibility by satisfying the principle of alternate possibilities. Williams' distinction reveals that this is not the notion of responsibility under discussion. The doubts thereby raised about these interpretations are confirmed when we turn to the details. I limit myself to three points—two on the relative surface of the passage, and one philosophically richer.

First, on a strictly textual level, the skeptics insist that the standard interpretation misreads Nietzsche's German. Complaints have centered on two terms—the reference to "morality" as '*Sittlichkeit*,' not '*Moralität*,' and Nietzsche's use of the modal verb '*dürfen*' to express the "right," or "permission," to promise. I rather doubt that the distinction between '*Sittlichkeit*' and '*Moralität*' can be made to bear Hegel-like weight across the board in Nietzsche's usage,* but the critics are correct that in this passage '*Sittlichkeit*' *is* specifically attached to the "morality of custom," and thus, when Nietzsche calls the sovereign individual "autonomous and super-moral [*übersittliche*]," the character is thereby placed outside the morality of custom in particular, not outside morality altogether. It does not follow, however, (as Hatab, *et al*., would have it) that the sovereign individual and his autonomy should then be understood *within* the framework of modern altruistic morality (understood as the successor to the morality of custom).[19] The form of morality (or political theory) Hatab would like to find in the passage would be a paradigm case of the altruistic "good/evil" value pattern Nietzsche criticized in the *Genealogy*'s First Treatise, whereas the sovereign individual's value scheme quite clearly falls under the opposed "good/bad" pattern Nietzsche ascribes to noble types. Rather than assuming a single moral standard apt for everyone, as altruistic morality would, the "responsibility" at the heart of his distinctive character is explicitly understood as a "privilege,"

and it is supposed to underwrite a "*standard of value*" based on the rare form of self-mastery he shares with a few other "strong and reliable" individuals, which licenses treating all others with "contempt" because they lack that kind of character (*GM* II, 2).[20]

Acampora (2006, 148-9) insists that the standard reading misinterprets the modal verb '*dürfen*'—the operative term describing the individual's "right," or "permission," to promise (he is someone "*der versprechen darf*"; literally, "who *may* promise"). She downplays the deontic modal sense that '*dürfen*' usually carries ('may' in the sense of allowance or permission) in favor of a sense of the verb emphasizing capacity (as in "One may well ask whether…," etc.). On this basis, she suggests that we are wrong to hear the sovereign individual's responsibility as any desirable "entitlement" (Acampora 2006, 148, 149); instead, it is a *mere* capacity, rooted in the power of memory and standing opposed to a countervailing capacity of forgetting. In my view, this is a stretch. The deontic sense of the modal verb is the normal one, and it seems needed here to underwrite Nietzsche's talk of responsibility as a "privilege" and a "*standard of value*," supporting esteem that the individual "'*earns*,'" or "'*deserves*'" (*verdient*) (*GM* II, 2). While *GM* II, 1-3 does describe the formation of a capacity for memory through a purely causal program of pain-based discipline and conditioning, that training is presented *merely as an instrumental means*—a "condition and preparation" (*GM* II, 2) and not a sufficient realization—for the permission to promise.[21] Indeed, that "preparation" is already present in (and is "the enormous work of," *GM* II, 2) the morality of custom which the sovereign individual transcends. Nietzsche's deontic modal '*dürfen*' thus points forward toward a normative achievement in the sovereign individual (a *conscience*), which is built on top of, but goes beyond, the capacity of

memory. In sum, the ground-level textual points marshaled to support the skeptical readings cannot support the weight rested on them.

Second, I also want to concede to critics like Leiter and Acampora that Nietzsche's genealogy of memory is intended to be, and is, a thoroughly naturalistic account of its emergence as a straightforward psychological capacity, even if it also serves as a basis for subsequent normative achievements like the sovereign individual's conscience (and eventually, the moralized concept of guilt).[22] Tellingly, however, the main consequence Leiter aims to derive from this naturalism is explicitly belied by the text itself. In the service of his argument that no substantive notion of autonomy could be at work in Nietzsche, Leiter is keen to promote a broad rejection of any causal role for conscious states within psychological life; if all apparent causation via conscious thoughts were epiphenomenal, then there could be nothing like self-determination via causal processes running through the agent's mind or will, and so, no such thing as autonomy in the compatibilist sense. As Leiter puts it,

> Denial of the causality of "the will" (more precisely, what we *experience* as willing) is central to Nietzsche's skepticism about free will (Leiter 2007)… If the faculty of the will "no longer 'acts' or 'moves'" (*A* 14)—if it is no longer causal—then there remains no conceptual space for the compatibilist idea that the right kind of causal determination of the will is compatible with responsibility for our actions. [Leiter 2011, 104]

Not two pages later, however, we find Leiter implicitly assuming that the very naturalist account he emphasizes (tracing the role of memory training as a basis for the permission to promise) involves a denial of any global epiphenomenalism about consciousness:[23] "Memory is essential

[to promising] for the obvious reason that only someone who can *remember* his promises can possibly honor them" (Leiter 2011, 106; emph. orig.). But why? Why wouldn't it be just as good for the promisor to form a wholly automatic disposition to act at the later time in a way suitably correlated with the promised performance, but without any memory of the promise itself? Memory is relevant, I submit, because the later action *counts* as performance (or not) only if the agent is *aware* (in a way that prompts her act) that she promised and how the present action relates to that promise, contrary to Leiter's official epiphenomenalism.[24]

It is worth re-emphasizing that the basic psychological capacity of memory whose history is being traced here is *only* a "condition and preparation" (*GM* II, 2) for the permission to promise that is Nietzsche's quarry. For it turns out that the fuller mechanism underwriting that permission is not just the cognitive ability to recall information but a more ambitious practical capacity involving genuine causation through the will, whose operation is deeply incompatible with the skeptical readings. With this, we arrive at the third and philosophically richer point about Nietzsche's portrait of the sovereign individual.

The skeptical readings all start from the thought that the sovereign individual remains caught in (standard, modern, bourgeois) morality; hence (given Nietzsche's "immoralism"), we should disregard the apparent praise built into his description. The most crucial support for this judgment is that the sovereign individual is a *promisor*, and the significance of promising is moral—indeed, deeply rooted in the conceptual and evaluative presuppositions of standard morality. The conclusion is natural to anyone thinking about promising in conventional terms, and even more natural to readers of the recent promising literature, which focuses somewhat

obsessively on worries about whether a promisor will deliver and what normative bond the promisee could rely on to guarantee performance (see Dannenberg 2015, 180). Those considerations *do* belong entirely to the domain of interpersonal morality—more particularly, to the domain of holding responsible, in the sense of Williams' distinction.

But Nietzsche's talk about responsibility in *GM* II, 1-3 is unconcerned about holding responsible, and in that light, it suddenly becomes clear how deeply *unconventional* his story is, as an account of promising.[25] It fails even to make mention of the promisee, or the promisor's obligation. Instead, the leading idea is the unusual notion that the right to promise depends on what Nietzsche calls "a true *memory of the will*" (*GM* II, 1). The appeal is not to cognitive recall, but instead to a fundamentally practical, will-based *analogue* to ordinary memory. Nietzsche starts from the supposition that humans have a strong and important power of forgetting, which is not only hard to resist but essential to our self-regulation.[26] It therefore requires a remarkable counterforce to overcome normal forgetfulness and make a promise:

> 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 once willed, a true *memory of the will*: so that a world of strange new 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 will breaking. [*GM* II, 1]

The persisting psychological structure sketched in this passage does presuppose the capacity of ordinary memory, as Nietzsche goes on to note two sections later, but the key persisting state is practical, not cognitive. It is "a *willing* on and on," a "long chain *of will*" (my ital.); it consists in the continuity of a distinctively practical attitude that commits the agent to an action. That is why intervening "acts of the will" pose a greater *prima facie* challenge to its psychological continuity—(Nietzsche remarks that "even" new acts of will can intervene)—than other "things, circumstances" that would be represented by strictly cognitive states in the agent.[27] Thus, what matters to his account of promising is that the agent's practical commitment itself persist, despite its not remaining in consciousness as an occurrent state. Cognitive memory and its making are part of the preparation, the means, but they are still the harsh, sour, unripe fruit—practical memory, practical reliability, "a true *memory of the will*" is the goal, the ripe fruit.[28]

So the heart of Nietzsche's intervention is this curious, non-occurrent but persistent and controlling, practical commitment, indicated through the metaphor of a memory belonging to the will itself. Jorah Dannenberg (2015) has recently offered a penetrating philosophical analysis of what such "memory of the will" might come to, and why it is important to promising. He understands memory of the will as a condition of continuity or stability in a person's *values*, resulting from her *active commitment* to create and/or support that very continuity. The problem that such a memory is suited to solve is underappreciated in theories of promising, although familiar enough in real life—it is a special form of *weakness* of the will whose manifestation is

not action against my considered judgment, so much as fatigue, or atrophy, or a failure of endurance in my commitment to what I value. As Williams would have been the first to point out, our valuing in real life cannot rest on reflective rational deliberation alone, but must be supported and partly constituted by a vast array of emotions, implicit judgments, and other immediate evaluative reactions to various stimuli thrown at us by the world.[29] It is an important feature of valuing that these reactions evolve in response to the stimuli, helping us to learn, and thereby to improve our fit with the changing environment, our effectiveness, and our satisfaction with our lives. But that very evolutionary process potentially threatens commitments to the things we value most. By "taking them for granted," by simply not attending to those things enough and so not nourishing the evaluative reactions through which we honor them, we can gradually lose the profile of immediate and emotional reactions through which we are attuned to them, and our valuing is threatened with atrophy. A special class of *promises* is distinctively oriented toward arresting such atrophy and thus toward strengthening our values. Dannenberg's (2015, 164-71) primary argument in the paper is that this is the best way to understand the widespread phenomenon of making promises *to yourself*, which he rightly argues would otherwise be deeply mysterious.[30] But once you apprehend the problem space, it is clear that certain kinds of interpersonal promises are in fact mainly about reinforcing memory of the will in this sense, rather than about guaranteeing performance. My promise in marriage to love and stand by and celebrate my wife did not aim to reassure her about my performance through creating an obligation to which she could hold me in case the architecture of my emotional orientation toward her eroded, and I was no longer inclined to be with her. If that state of my

psychology developed, the marriage as we conceived it would already be in trouble and my promise already violated. On the contrary, the promise was a commitment to her—and simultaneously to myself—to strive not to let any such drift of my emotional orientation happen in the first place. It was a pledge to create and conserve a memory of the will capable of sustaining my commitment to her. Anyone without the capacity for that special sort of memory—(Nietzsche's phrasing suddenly becomes strikingly apt)—would not really have the *right* to make that kind of promise.

Dannenberg (2015, 172-8) shows that this capacity also has more to do with interpersonal promises in general than people have appreciated. What is at stake in promising to yourself, and in those promises like marriage or friendship where one's basic identity—one's mature agency—is on the line, is the ability to *bind yourself* at all, what Nietzsche calls the ability to "vouch for [yourself] *as future*" (*GM* II, 1), as someone capable of sustaining your values. Even though contractual promises are not really Nietzsche's concern in the passage, such self-binding is also involved in those conventional cases, in that the bond I put on *myself* is crucial to performance in the right (or anyway, in the best) spirit. For example, we might hold with Dannenberg (2015, 177-8) that when I promise and induces your reliance, I commit myself to valuing your point of view when the time comes to perform, and thus to memory of the will regarding that value.

But there remains an important difference in the basic moral analysis of these two types of promise, parallel to Williams' distinction. In contract-like, standard interpersonal promises, my capacity to bind myself is deployed as one piece within a larger moral structure that is governed by holding responsible; I undertake an obligation to you, you expect performance, and

then even if I perform reluctantly (only in recognition of the obligation to which you hold me), it still counts as (minimal) performance. By contrast, in the case of promises to myself—and likewise in deeply personal promises like those of friendship or marriage that commit not my external acts but my mature agency itself—this wider structure is absent, and holding responsible is out of place. Instead, the normative stakes all along are *just* self-binding, i.e., taking responsibility in the sense that touches the overall shape of my life. Either I will succeed to sustain my commitments in that sense or not, but either way, the attempt to hold me responsible from the outside by interpersonal obligation must misfire—if I sustained the commitment to my values, then compliance was there already; and if not, then any compliance that has to be secured externally by holding me responsible will already have let my wholehearted self slip through its fingers. Even if my friend is inclined to *judge me wanting* against the standard of mature agency when I fail to perform, this is not helpfully seen as holding responsible. As soon as she must appeal to the language of compensation and guilt proper to impersonal moral accountability—as opposed to calling for Williams-style agent regret and renewed commitment to our relationship from my side, and expressing concern for the shape of my life from hers—then it becomes clear that my betrayal was too great and the real friendship is over. Such talk is no sign that my promise all along operated in the space of holding responsible, but a sad acknowledgement of our relationship's deterioration to the impersonal level. The deeply interpersonal promises of close friendship really do belong to a different normative domain than impersonal contracts.[31]

With this, we can see the deep sense in which the skeptical readings are misguided. They hear Nietzsche's talk about the sovereign individual's responsibility and his promises entirely in

the spirit of "contract," "commercial virtues," and the promisee's demand that the promisor be held responsible to meet her obligations. But this mistakes the register of Nietzsche's concerns about promising, which have to do with the individual's ability, *and right*, to take responsibility for herself in a way that organizes and unifies her life as a whole, and not with holding responsible at all. The underlying philosophical motivations of the skeptical readings—rejection of unified selfhood, the assumption that appeals to autonomy must be meant to underwrite moral responsibility—are likewise off the mark. In the light cast by Williams' distinction, Nietzsche's intentions appear almost diametrically opposed to the ones found by the skeptical readers. Where they seek a disunified, plural self, Nietzsche is concerned with how an individual can unify herself around her values; where they see external, contractual promises among anonymous bourgeois traders, he is concerned with the promises a person makes to herself, or ones that put her whole self, the intimate self engaged by her friends, on the line; where they see the sendup of contra-causal or alternate-possibility freedom, Nietzsche is concerned with autonomy as a structural feature of character capable of making a person's life her own by manifesting not only her right to take responsibility for her self and her acts, but also her entitlement to genuine satisfaction with them. "To be permitted to vouch for oneself, and with pride, hence to be *permitted to say 'yes'* to oneself too—that is, as noted, a ripe fruit" (*GM* II, 3).

We can now return to questions about autonomy. The sovereign individual was supposed to be *autonomous* as well as being (or, because he is) responsible. But what must Nietzsche mean by 'autonomy,' if the Williams insight fixes the relevant kind of responsibility?

4. Nietzschean Autonomy

Most immediately, the sovereign individual's autonomy manifests as *independence*—he is *free from* the morality of custom. That initial point is marked by Nietzsche's parenthetical remark that "autonomous" and "moral" exclude one another (*GM* II, 2). But it is *obvious* that individual autonomy is incompatible with the unquestioned authority of custom—*so* obvious that no remark would have been needed if that were Nietzsche's only point. The parenthetical must be intended to evoke some deeper tension between autonomy and morality, going beyond morality in the sense of custom. The effect is to cast shade on the tight connection between the two notions assumed by Kantian conceptions of moral autonomy,[32] but it also raises questions for the reader about why the sovereign individual would depart from the safety of his society and morality in the first place. In answer, Nietzsche offers no hint that the individual could have been "liberated" by some external change in his concrete social or political relations toward those who remain in the sway of the morality of custom. Instead, what changes is the *internal* structure of his psychology. He develops a "*memory of the will*" that enables him to bind himself and stabilize his values, cultivating a "mastery over himself" that allows him to "vouch for himself" and eventually to "*say 'yes'* to himself" (*GM* II, 2, 3). Thus, he no longer needs the constraints of the morality of custom nor the harsh threats of violent punishment to enforce the basic norms that permit social life (*GM* II, 3). Rather than being bound and controlled by external threat, the individual binds and controls himself. Independence from society and morality thereby turns out to rest on autonomy in a deeper sense of self-control or self-governance. At this deeper level, the sovereign individual's autonomy is neither a suspension of

external social constraint nor a power to have acted otherwise, but a certain *self-relation*, a structural property of her character.

What kind of self-relation is this, and what justifies Nietzsche's description of it as "autonomous"? Williams' distinction suggests two morals bearing on the question. From the side of mature agency, the suggestion is that this structural property of character finds its home within an *ideal-bearing* account of the self. Thus, it will be aspirational, demanding to attain, and achieved as a matter of degree. From the other side, we saw that the notions of self-control and holding responsible used for social coordination claim a different home terrain from this ideal-bearing self-relation, leading to systematic differences of application. The demanding ideal may be realized by only a narrow domain of people, while for those who attain it, it must encompass their life as a whole (or something close to it). By contrast, the sort of self-control linked to holding people responsible, and so suitable for duty within standard compatibilism, must apply to a very wide domain of people (all those whose behavior is to be regulated by the social mechanisms of accountability), while it correspondingly covers a narrower part of their lives—just what makes a necessary claim on the attention of the public, as Williams put it. This gives us a distinction between the sort of narrow-domain, aspirational autonomy that was Nietzsche's concern, and the wide-domain form of autonomous self-control that might be of use for the modern compatibilist project.

A third feature for the picture of Nietzschean autonomy can be found in a recent paper by Donald Rutherford (2011). He reminds us that for many in the historical tradition, the key mark of freedom was not alternate possibilities for action but action in accordance with reason. In

early modern philosophy, the latter idea was associated with an "intellectualist" tradition, which defended the practical priority of intellectual representations of the good against voluntaristic positions committed to liberty of indifference, but Rutherford emphasizes that such accounts were also attractive to frankly necessitarian philosophers like the Stoics and Spinoza, which offers suggestive possibilities for interpreting Nietzsche. As he concedes (Rutherford 2011, 521), there are bound to be major differences between any Spinozistic view and Nietzschean autonomy. For unlike Nietzsche, a Spinozist would accept the intellectualist identification of autonomous activity with the operations of reason.[33] Adhering to my power of reason is supposed to liberate me from the operations of passion that render me vulnerable to other things as a causal patient. Since Nietzsche is inclined to identify the "self" involved in *self*-governance with exactly such non-rational drives, affects, or passions, the *content* of the Spinozistic solution is not available to him. Rutherford's intriguing point, however, is that the *form* of the strategy does remain available. What makes for Spinozistic agency, as opposed to passivity, is determination by internal rather than external factors, and while Spinozists spell out the internal in terms of the essential power of reason, a Nietzschean alternative might start from the same structure, while substituting a different conception of what is "internal" to the agent.

Rutherford's preferred version of the strategy identifies the Nietzschean self with a fixed hierarchy of drives controlled by a dominating instinct (Rutherford 2011, 528-9). Such a person would be autonomous if her actions were dictated by the internal configuration of drives that is essential to her, and unfree if they were determined by external causes. From the point of view opened up by Williams' insight, this account remains a bit too focused on delimitation of a

special set of (quasi-)voluntary actions, but in any case, I am skeptical that it can be made to fit Nietzsche's other commitments. The reasons why expose philosophical problems faced by any adequate conception of Nietzschean autonomy.

To be clear, Nietzsche does often indulge in talk of an individual's "dominant drive,"[34] but I doubt that such talk ever represents commitment to a psychological essence,[35] as opposed to a provisional (or occasionally, a degenerated) state of affairs in the person's psychology.[36] His picture of the "drive dynamics" within the soul involves too much change and instability to allow any fixed drive hierarchy a credible claim to speak for the "genuine self" across all changes. As Hatab and Acampora emphasize, Nietzsche is just as willing to praise multiplicity and tensions within the self as he is to esteem its unity,[37] and where he does advocate unity, he tends to represent it as something *achieved* through the harmonization of competing drives, not as a pre-given structure in which one drive dominates (Katsafanas 2011, Anderson 2012). Indeed—and here the real problem surfaces—Nietzsche's underlying commitments threaten to tell against the recognition of any fixed boundaries separating "the self" from "the other." What is distinctive and (frankly) <u>weird</u> about the units of will to power that Nietzsche wants to posit at the basis of things is that their activity makes everything a potential target for incorporation[38]; they are "trying to eat the world," as Elijah Millgram once memorably put it to me.[39] So the philosophical problem is not just that Nietzsche rejects any fixed drive structure capturing the essence of the self, but that he may deny himself the resources to mark off the internal from the external at all, in the way our strategy requires to separate autonomous from alienated activity.[40]

One further complication deserves separate mention. Among the texts that suggesting

Nietzsche's commitment to autonomy as a value, many are strongly inflected by the idea that *the agent herself can shape her drives*, reforming her original nature into a new, *second* nature through activities of self-training (see esp. *GS* 290, also 299, 335; *UM* IV, 11; *BGE* 203). That would offer a further sense in which the agent takes responsibility for herself, realizing an autonomy like the sovereign individual's. But if the agent can *change* the structure of her drives, and if the new configuration counts as more her own (more autonomous) because of that history, then we cannot legitimately appeal to any *given* structure of drives to separate the internally driven from the externally imposed activities.

Williams' emphasis on the *ideal-bearing* character of taking responsibility for yourself again becomes helpful at this juncture. On such an ideal-bearing account, responsibility (and ultimately, selfhood itself) are always attained only to some degree. That fits better with the Nietzschean themes just canvassed—the idea that the self can pursue such an ideal by working on a second nature for itself; the idea of the self as a collection of drives, affects, and other attitudes with some imperfect degree of integration; the idea that each self (and each drive within it, etc.) is always aiming to incorporate more of the environment, and in that sense refuses to recognize any fixed boundary. Perhaps, then, the idea of *approximation* to ideal selfhood might offer a way to identify a *non-fixed* internal/external boundary that is determined provisionally and contextually, depending on the degree of approximation to the ideal?

Another thing we learned from Williams' distinction is that Nietzsche's thinking about freedom was never trying to save the principle of alternate possibilities, or more broadly, the "wide domain" compatibilist ambition to find some conception of control applicable to the full

range of people one might want to hold responsible. Nietzsche thought such efforts were not just doomed but pointless—that they were rooted in altogether *the wrong question* about freedom.[41] One core text making that case is *BGE* 21, where Nietzsche roundly rejects *both* the doctrine of "free will" and also "unfree will" (conceived via negation of "free will"—e.g., through determinism). He then tries to change the subject away from the traditional question of free will with this cryptic sentence: "The 'unfree will' is mythology; in real life it is only a matter of *strong* and *weak* wills."[42]

As I read it, the suggestion is that the real problem about freedom has to do with an ethical problem of weakness of will, construed broadly to include not only standard akratic action against one's best judgment, and the sort of atrophy of conative orientation identified by Dannenberg (2015), but even more—an ethical problem of weakness that is both broader and more widespread than we commonly think. It is broader in that the basic psychology of *inner conflict* behind standard-issue akrasia has analogues in many cases where we are not akratic, strictly speaking. For example, the *ressentiment*-motivated priests of *GM* I need not be akratic, since they may act in accordance with their official, slave-morality-inspired values, but the persistence of vengefulness in their psychologies indicates a deep-going evaluative inconsistency that points toward a basically self-defeating overall pattern of valuation (Reginster 1997). In Nietzsche's mind, such cases of inner conflict deserve to be counted as ethico-psychological weakness by extension, because of their analogy to the kind of inner conflict among desires and values that generates standard akrasia. Weakness is also more widespread than we think, in that *very* many ordinary actions (including even clear instances of intentional, voluntary agency) are

really weak-willed in the extended sense even if we do not know it, because we are merely "following along" in the activities and pathways suggested for us by our social and motivational environment, rather than forging a path for ourselves out of true independence of spirit (Gemes 2009). Nietzsche's diagnosis is that we "follow along" precisely because internal divisions leave us vulnerable to alienating influence from others. As I have argued elsewhere (Anderson 2006, 2012), the ubiquitous talk of strength and weakness throughout Nietzsche's writing is fruitfully interpreted in this spirit, as a matter of ethical weakness of the will, broadly construed.[43]

Nietzsche's own account of the phenomenon lines it up with a defensible separation of what is internal to the self from what is external. His definition of weakness of the will is "the inability *not* to respond to a stimulus" (*TI* V, 2)—that is, the susceptibility to having one's activities dictated by (external) environmental stimuli. Such activity is alienated, hence unfree. One is *unable* not to respond, Nietzsche thinks, precisely when and because the drive receptive to the stimulus is insufficiently integrated with others in the self. Nietzsche also endorses a corresponding notion of strength of will, which amounts to a capacity for self-control: "strong will: the essential feature is precisely *not* to will—to *be able* to suspend decision" (*TI* VIII, 6) in the face of external inducements. On this reading, Nietzsche's claims about weakness and strength are not an embarrassing appeal to brute, physical characters of size, strength, and violent behavior, but instead have a moral psychological basis clearly related to autonomy. Weakness amounts to a 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 control by

the whole self. 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 when some drive is demanding it. In short, strength and weakness are a matter of unity and disunity among the constituent drives and other attitudes of the self.[44] This feature is attained as a matter of degree and is plausibly related to an ideal of internal coherence, so it can be understood as a Williams-style ideal-bearing account.

But how do we mark the *boundaries* needed to underwrite the contrast between what is "internal" to the person and the "external" stimuli impinging upon her? As we saw, Nietzsche denies that the self is given within fixed boundaries, but one strand of his skepticism about this point connects helpfully to the present account of strength and weakness. Nietzsche insists on "*anti-atomism*" about the soul: "Let it be permitted to designate by this expression ["*soul atomism*"] the belief which regards the soul as something indestructible, eternal, indivisible, as a monad, an *atomon*: This belief ought to be expelled from science!" (*BGE* 12). In my view, this anti-atomism is supposed to be quite radical indeed[45]; it denies not only the traditional predicates of substantiality, simplicity, and immortality, but also permanent boundaries (if the soul is a "social structure of drives and affects" (*BGE* 12), then it might gain or lose "constituents"; see also *WP* 488 = *KSA* 12: 391-2), and even, supposing atomism is really to be "expelled from science," also atomism about *parts* of the self. If there are no atoms to be had, then a drive (or other self-constituent) must lack simplicity just as much as the ego itself, and the condition of internal complexity must go "all the way down," each drive being composed of further drives, and so on. Appeal to a conception of strength as *degree of unification* helps us understand both

how such an anti-atomism could possibly be coherent, and also how we can provisionally distinguish what is internal to some self (or constituent) from what is external. Nietzsche's idea must be that every psychological entity has some degree of unity or internal order. Insofar as it is sufficiently unified for its constituents to act together, it counts as one thing, with an internal domain of constituents (this is strength); but given anti-atomism, it always *has* constituents and is therefore potentially subject to weakness manifested as inner division, which, if it progressed far enough, would yield disintegration. In its efforts to "eat the world," each such entity strives to incorporate others, moving them from the "outside" to the "inside" by integrating their tendencies of behavior into its own set of more or less coordinated parts. If successful, it manifests greater strength by exerting control over a greater variety of elements, but it proceeds always at its own risk, since the process may simply lower the overall level of integration, resulting in weakness (or even disintegration). The boundaries of the self are variable just because, and just insofar as, it is vulnerable to weakness, but weakness is always a matter of degree, so whether an internal/external boundary can be fixed and just where it belongs will be contextual matters depending on the overall circumstances of the relevant interaction(s).[46]

In general, then, Nietzschean strength is a matter of the *integration* of the self's drives and other attitudes so that they cohere to form an individual. In the strong self, the integrating order (which I take to *be* the self[47]) settles the place of component drives within it, and exploits their tendencies for its larger ends. This commanding self thereby identifies with the drives and their activities, just as in a successful commonwealth the governing class identifies with the whole (*BGE* 19). It represents the drives (along with their "triumph over obstacles" (*BGE* 19))

as belonging to *it*, and endorsed by it. In that sense, the well ordered, self-consistent, fully individual person acts from internal principles; she is autonomous.

Nietzsche's views about autonomy thus turn out to be illuminated by appeal to his psychology of strength and weakness, just as the cryptic remark from *BGE* 21 suggested. Failures of autonomy are cases of psychological weakness in the extended sense, traceable to inner conflict among the agent's attitudes in which the whole is unable to control some recalcitrant drive. The achievement of autonomy, by contrast, is the distinctive self-relation of strength—a coherent integration of the attitudes, in which each is governed by its place in the whole. The key texts where Nietzsche praises freedom or autonomy rely on just such ideas: for example, he describes the achievement of Goethe (who has "*become free*") this way,

> What he wanted was *totality*; he fought the mutual extraneousness of reason, senses, feeling, and will (preached with the most abhorrent scholasticism by *Kant*, the antipode of Goethe); he disciplined himself to wholeness, he *created* himself. [*TI*, IX, 49; cp. *GS* 335, 347]

Of course, as Nietzsche was all too well aware, in real life we are routinely plagued by more or less serious forms of internal conflict, partly evoked by the pressure worldly obstacles mount against the joint satisfaction of our various desires, but partly owing simply to imperfect integration of our own attitudes and commitments themselves. In this sense, the ideal of Nietzschean autonomy is aspirational and demanding, in the ways Williams detected in ideal-bearing accounts of responsibility. It will be realized to a serious degree by a relatively narrow domain of people, and it ambitiously pretends to extend to a very wide swath of their lives. This confirms that it is not the sort of "wide domain" autonomous self-control apt for the compatibilist

project of vindicating our efforts to hold one another responsible. But it is a remarkably good fit for that sovereign individual from whom we started. His efforts to extend his responsibility give him a mastery over himself that unifies him and makes him "*strong* and reliable," "*strong* enough to uphold [his will] even against accidents, even 'against fate'" (*GM* II, 2; my ital.). More, his taking responsibility, his striving to ensure the stability and continuity of his values into the future, becomes a key mechanism contributing to the integration of his attitudes, and thus to his strength *sensu* Nietzsche: developing "memory of the will" is the *means* by which he achieves his approximation to the ideal of autonomy. Like Goethe, he has "become free" (*GM* II, 2; cp. *TI* IX, 49)—and he did so in large measure through his own efforts to take responsibility for himself. In the end, his autonomy and his mature agency are deeply intertwined.

5. Conclusion: Lessons from Nietzschean Autonomy

A good deal of the picture I sketched while exploring Nietzschean autonomy rests on aspects of his wider philosophical vision, some of which perhaps strike us as idiosyncratic. Other elements, though, operate at a sufficiently abstract level, and are thought through in a sufficiently determined way, that they offer instructive materials for philosophical development. In closing, I would like to highlight five such features. Two are interconnected ideas that Nietzsche shares with Williams, one is proper to Nietzsche, and two might be derived from thinking in Williams' spirit *against* Nietzsche, at least in some of his moods.

Turning first to the common ideas, Williams articulates an important insight when he insists that the ideal of mature agency attained by taking responsibility for yourself is to be

distinguished from, rather than deployed as the basis for, the practice of holding one another responsible in social relations. The same (or a very similar) idea turned out to be involved in Nietzsche's real philosophical point about promising from the beginning of *Genealogy*, II. The ability to bind oneself through promises that put one's mature agency at stake turns out to be important because of its tendency to form and strengthen one's character (i.e., one's mature agency itself), and not because of any guarantee to a promisee or any interpersonal obligation.

On the related second point, Williams' distinction can serve not only to illuminate what is at stake in the vexed text about the sovereign individual, but also to clarify different philosophical uses we have for the idea of autonomy. The sovereign individual's autonomy is to be understood as a self-relation pertaining to one's character and bound up with a remarkable ability to stick to one's commitments. That ideal is not at all easy to attain, if we honestly take account of the vulnerability of our valuing to obstacles from the side of the world and weakness from the side of ourselves. We are wrong to demand approximation to the ideal when we seek a notion of autonomous control of actions suitable for holding one another responsible in social life. If such a project owns up to the ideal's demandingness, it will let too many agents off the hook as if they were irresponsible wantons, rather than what they are, which is normal inhabitants of the vulnerable human condition. If we don't own up, then we drain the ideal of mature agency of a key part of its power to inspire, which derives in no small measure from the demand it makes on us to take responsibility for things that we could never fairly be held responsible for, thereby integrating them into our lives in a real and meaningful way.

The idea from Nietzsche I want to recall is abstract. We can make progress in

understanding the ideal of responsible mature agency by leaning on the idea of what is determined by, and hence *expresses*, that which is *internal* to the self—understanding the "internal" as what *unifies* (more or less, although often less) the person's complex, internally contested psychology, rather than through some a priori conception identifying the true self with reason, or the like.[48] When my life is more in accord with what is internal (what most unifies it), then I will better approximate the ideal self-relation of genuine interest, showing some progress toward taking responsibility for myself. And that ideal *is* of genuine interest: when Williams remarks that it is not a sign of maturity to take responsibility *only* when impersonal morality holds us answerable, he reminds us how much there is about the overall shape of our life that genuinely matters, but which resides beyond or below that sphere dominated by reason, or by impersonal relations and our efforts to hold one another responsible.

But it is not as though these two different spheres are entirely unrelated, and we can understand their relations better by turning to Williams for a fourth point. *Both* holding one another responsible *and* taking responsibility for ourselves are centrally important to the value of our lives overall. For that very reason, the two forms of responsibility are bound to interact *at the level described by value theory*, operating to reinforce or to interfere with one another when we come to all-things-considered assessment of a life. Thus, on the side of reinforcement, we saw that my honoring my impersonal obligations can acquire enhanced value from being carried out "in the right spirit"—where that means in a spirit that fully engages my mature agency—and conversely, it can contribute to the value of intimate friendship that the friends hold one another to their promises and ideals. On the side of interference, recall that one of the morals of

Williams' famous Gaugin example was to show the possibility that values rooted in the shape of one's life overall could come into substantial, irreducibly evaluative conflict with values rooted in obligations to which others would like to hold us responsible. Part of what I have meant to suggest today is that once we leave off trying to explain these relations of support and interference by identifying some point within the philosophy of action that could connect the two kinds of responsibility in terms of a deeper theory about what responsibility *tout court* metaphysically *is*, we can get on with the messier, but ultimately more promising, project of mapping their different and interacting contributions to our lives by assessing just why and just how they actually matter to the concrete value of a given life.

In one last respect, finally, it seems to me that the spirit of Williams' thought has more to teach us than the tone of some of Nietzsche's own remarks about autonomy. In his enthusiasm to dramatize the alienating threat of social conformity, Nietzsche sometimes writes as if *real* autonomy were a matter of achieving a purely individual way of life, hermetically insulated from the taint of worldly connection and social relations. In that spirit, for example, he describes the sovereign individual as someone "who promises like a sovereign, weightily, seldom, slowly, who is stingy with his trust," and who "reserves his kick for the dogs who promise although they are not permitted to do so" (*GM* II, 2)—that is, one fears, for any normal person at all subject to vulnerability in valuing. Or again, the strong-willed individual is "separated from the crowd and its duties and virtues," full of "loftiness of glances that dominate and look down," "the pleasure and exercise of the great justice, the art of command, the width of the will, the slow eye that rarely admires, rarely looks up, rarely loves" (*BGE* 213). In passages like these, Nietzsche

succumbs to a tempting misconstrual of the ideal-bearing form of autonomy, which overreads its proper *independence* from external determination as if it required thoroughgoing *isolation* from all meaningful interactions whatsoever. But this is just a mistake, and on his own premises Nietzsche was never entitled to it. After all, once we concede that the internal/external boundary itself is variable and only contextually fixed, and that (another perfectly Nietzschean insight[49]) the meaningfulness of actions that permits them to express one's internal self at all is thoroughly dependent on their interpretability in a social context, then it becomes clear that such isolation is neither possible nor desirable. Autonomy is not about cutting yourself off from the social world, but inserting your genuine self (and its internal principles) into it. Williams, with his relentless resistance against the Stoic-style program of investing all value in some retrenched domain of "what is purely up to me"—impregnably defended against the basic vulnerability to luck built into our finite human condition—is well positioned to offer reminders to Nietzsche on this point. What matters about our autonomous mature agency is not how well we can protect it, but that we can actually use it in the world—to act for ourselves and what we value, but also just as important, to take responsibility for ourselves, not only in our voluntary acts but in our vulnerable finitude.

References

Works by Nietzsche

For Nietzsche's German, I used Nietzsche 1980 ff. (*KSA*). I also made use of the following translations, cited by abbreviations, although I occasionally make minor departures from their renderings without separate notice. Parenthetical citations in the text refer to Nietzsche's section numbers, which are the same in all editions.

KSA Nietzsche, F.W. (1980 ff.) *Werke: Kritische Studienausgabe*, ed. G. Colli and M. Montinari. Berlin: Walter de Gruyter.

UM Untimely Meditations, trans. R.J. Hollingdale. Cambridge: Cambridge University Press, 1983 (1873-6).

HH Human, All-too-Human, trans., R.J. Hollingdale. Cambridge: Cambridge University Press, 1986 (1878-9).

D Daybreak, trans. R.J. Hollingdale. Cambridge: Cambridge University Press, 1982 (1881).

GS The Gay Science, trans. W. Kaufmann. New York: Vintage, 1974 (1882, 1887).

BGE Beyond Good and Evil, trans. W. Kaufmann. New York: Vintage, 1966 (1886).

GM On the Genealogy of Morality, trans. M. Clark and A. Swensen. Indianapolis: Hackett, 1998 (1887).

TI Twilight of the Idols, trans. W. Kaufmann. New York: Viking, 1954 (1888).

CW The Case of Wagner, trans. W. Kaufmann. New York: Vintage, 1967 (1888).

A The Antichrist, trans. W. Kaufmann. New York: Viking, 1954 (1888).

EH Ecce Homo, trans. W. Kaufmann. New York: Random House, 1967 (1888).

WP The Will to Power, trans. W. Kaufmann and R.J. Hollingdale. New York: Vintage, 1967.

Works by other authors

Acampora, Christa. (2004) "On Sovereignty and Overhumanity: Why it Matters How We Read Nietzsche's *Genealogy* II, 2," *International Studies in Philosophy* 36: 127-45.

-----. (2006) "On Sovereignty and Overhumanity: Why it Matters How We Read Nietzsche's *Genealogy* II, 2." In Acampora, C., ed., *Nietzsche's On the Genealogy of Morals: Critical Essays*. Lanham, MD: Rowman and Littlefield, pp. 147-61.

Anderson, R. Lanier. (2006) "Nietzsche on Strength, Self-Knowledge, and Achieving Individuality," *International Studies in Philosophy* 38: 89-115.

-----. (2012) "What is a Nietzschean Self?" In Janaway and Robertson, eds. (2012), 202-35.

-----. (2013a) "Love and the Moral Psychology of the Hegelian Nietzsche: Comments on Robert Pippin, *Nietzsche: moraliste français*," *Journal of Nietzsche Studies* 44: 158-80.

-----. (2013b) "Nietzsche on Autonomy." In Richardson and Gemes, eds. (2013), 432-60.

Ansell-Pearson, Keith. (1991) "Nietzsche on Autonomy and Morality: the Challenge to Political Theory," *Political Studies* 39: 270-86.

Bratman, Michael. (2007) "Reflection, Planning, and Temporally Extended Agency." In Bratman, *Structures of Agency: Essays*. Oxford: Oxford University Press, 2007, pp. 21-46.

Clark, Maudemarie, and David Dudrick. (2009) "Nietzsche on the Will: an Analysis of BGE

19." In Gemes and May (2009), pp. 247-68.

Craig, Edward. (1990) *Knowledge and the State of Nature: an Essay in Conceptual Synthesis*. Oxford: Oxford University Press.

Dannenberg, Jorah. (2015) "Promising Ourselves, Promising Others," *Journal of Ethics* 19: 159-83.

Gardner, Sebastian. (2009) "Nietzsche, the Self, and the Disunity of Philosophical Reason." In Gemes and May, eds. (2009), pp. 1-31.

Gemes, Ken. (2009) "Nietzsche on Free Will, Autonomy, and the Sovereign Individual." In Gemes and May, eds. (2009), pp. 33-49.

Gemes, Ken, and Simon May, eds. (2009) *Nietzsche on Freedom and Autonomy*. Oxford: Oxford University Press.

Hatab, Lawrence J. (1995) *A Nietzschean Defense of Democracy: an Experiment in Postmodern Politics*. Chicago, IL: Open Court.

-----. (2008) *Nietzsche's On the Genealogy of Morality: an Introduction*. Cambridge: Cambridge University Press.

-----. (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.

Irwin, T.H. (1980) "Reason and Responsibility in Aristotle." In Rorty, A.O., ed., *Essays on Aristotle's Ethics*. Berkeley, CA: University of California Press, pp. 117-55.

Janaway, Christopher. (2007) *Beyond Selflessness: Reading Nietzsche's Genealogy*. Oxford:

Oxford University Press.

-----. (2009) "Autonomy, Affect, and the Self in Nietzsche's Project of Genealogy." In Gemes and May, eds. (2009).

Janaway, Christopher, and Simon Robertson, eds. (2012) *Nietzsche, Naturalism, and Normativity*. Oxford: Oxford University Press.

Katsafanas, Paul. (2011) "The Concept of Unified Agency in Nietzsche, Plato, and Schiller," *Journal of the History of Philosophy* 49: 87-113.

-----. (2013) *Agency and the Foundations of Ethics: Nietzschean Constitutivism*. Oxford: Oxford University Press.

Leiter, Brian. (2007) "Nietzsche's Theory of the Will," *Philosopher's Imprint* 7: 1-15. Reprinted in Gemes and May, eds. (2009).

-----. (2011) "Who is the 'sovereign individual'? Nietzsche on Freedom." In May, ed., (2011), pp. 101-19.

May, Simon, ed. (2011) *Nietzsche's On the Genealogy of Morality: a Critical Guide*. Cambridge: Cambridge University Press.

Nehamas, Alexander. (1985) *Nietzsche: Life as Literature*. Cambridge, MA: Harvard University Press.

Owen, David. (2002) "Equality, Democracy, and Self-Respect: Reflections on Nietzsche's Agonal Perfectionism," *Journal of Nietzsche Studies* 24: 113-31.

-----. (2007) *Nietzsche's Genealogy of Morality*. Montreal: McGill-Queen's University Press.

Pippin, Robert B. (2009) "How to Overcome Oneself: Nietzsche on Freedom." In Gemes and

May, eds. (2009), pp. 69-87.

-----. (2010) *Nietzsche, Psychology, and First Philosophy*. Chicago: University of Chicago Press.

Poellner, Peter. (1995) *Nietzsche and Metaphysics*, Oxford: Oxford University Press.

-----. (2009) "Nietzschean Freedom." In Gemes and May, eds. (2009), pp. 151-79.

Reginster, Bernard. (1997) "Nietzsche on *Ressentiment* and Valuation," *Philosophy and Phenomenological Research* 57: 281-305.

-----. (2011) "The Genealogy of Guilt." In May, ed. (2011), pp. 56-77.

Richardson, John. (1996) *Nietzsche's System*. Oxford: Oxford University Press.

-----. (2009) "Nietzsche's Freedoms." In Gemes and May, eds. (2009), pp. 127-49.

Richardson, John, and Ken Gemes, eds. (2013) *The Oxford Handbook of Nietzsche*. Oxford: Oxford University Press.

Ridley, Aaron. (1998) *Nietzsche's Conscience: Six Character Studies from the "Genealogy"*. Ithaca, NY: Cornell University Press.

-----. (2007) "Nietzsche on Art and Freedom," *European Journal of Philosophy* 15: 204-24.

-----. (2009) "Nietzsche's Intentions: What the Sovereign Individual Promises." In Gemes and May, eds. (2009), pp. 181-96.

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

Schacht, Richard. (1983) *Nietzsche*. London: Routledge.

-----. (2006) "Nietzsche and Individuality," *International Studies in Philosophy* 38: 131-51.

Stevenson, Charles. (1938) "Persuasive Definitions," *Mind* 47: 331-50.

Williams, Bernard. (1981) *Moral Luck: Philosophical Papers, 1973-1980*. Cambridge: Cambridge University Press.

-----. (1981 [1976]a) "Moral Luck." In Williams (1981), pp. 20-39.

-----. (1981 [1976]b) "Persons, Character, and Morality." In Williams (1981), pp. 1-19.

-----. (1993) *Shame and Necessity*. Berkeley, CA: University of California Press.

-----. (1995 [1989]) "Voluntary Acts and Responsible Agents." In Williams, *Making Sense of Humanity and Other Philosophical Papers, 1982-1993*. Cambridge: Cambridge University Press, pp. 22-34.

Notes

1. Acknowledgements… JD, DO, BR

2. Readings in this broad class are very widespread, so I will refer to them as the "standard reading." It is worth noting, however, that the general idea that the sovereign individual is a target of praise can be cashed out in very different ways. Some versions that may be of interest to readers of this paper include Schacht (1983, 294-6; 2006), Ridley (1998, 142-6; 2009), Anderson (2006), Janaway (2007, 116-23), Owen (2007, 96-102), and Pippin (2009, 80).

3. All citations to Nietzsche are given parenthetically in the text, and follow the abbreviation scheme detailed at the head of the references. For the most part, I have followed the translations listed there, although I occasionally depart from them in the interest of greater literalness without separate notice. (In some such cases, I provide Nietzsche's German in brackets.)

4. Hatab and Acampora are reacting against those (Keith Ansell-Pearson (1991) is a prominent target) who, in their view, assimilate Nietzsche's thought too closely to certain liberal pieties (e.g., about the value of autonomy). Despite our other disagreements, I should concede that I share some sympathy with their caution on this point.

5. Thus, for Hatab, the sovereign individual can rightfully "vouch for himself" (*GM* II, 3) in pretty much the sense that the individual subject of Locke's political theory has an original property right in himself (Hatab 2009, 177) and thus a moral right to himself and the products of his labor.

6. Full discussion is beyond the scope of this essay, but readers can get a sense for the flavor of this strand of texts in Nietzsche from the following sampling:

> Let it be permitted to designate by this expression ["the *soul atomism*"] the belief which regards the soul as something indestructible, eternal, indivisible, as a monad, an *atomon*: this belief ought to be expelled from science! … But the way is open for new versions and refinements of the soul-hypothesis; … "mortal soul," "soul as subjective multiplicity," "soul as social structure of the drives and affects"… [*BGE* 12]
>
> *It* thinks; but that this "it" is precisely the famous old "ego" is, to put it mildly, only a supposition, an assertion, and assuredly not an "immediate certainty"… [*BGE* 17]
>
> But there is no such substratum [the "doer"]; there is no 'being' behind doing, effecting, becoming; 'the doer' is simply fabricated into the doing—the doing is everything. [*GM* I, 13]
>
> To indulge the fable of 'unity,' 'soul,' 'person,' this we have forbidden: with such hypotheses one only covers up the problem. [*KSA* 11: 577]
>
> We enter a realm of crude fetishism when we summon before consciousness the basic presuppositions of the metaphysics of language... . Everywhere it sees a doer and doing; it believes in will as *the* cause; it believes in the ego, in the ego as being, in the ego as substance... that calamity of an error. [*TI* III, 5]
>
> And as for the ego! That has become a fable, a fiction, a play on words: it has altogether ceased to think, feel, or will! [*TI* VI, 3]
>
> We suppose that *intelligere* must be... something that stands essentially opposed to the instincts, while it is actually nothing but a *certain behavior of the instincts toward one*

another. [*GS* 333]

7. Leiter (2011, 101, 110-12) cites Ansell-Pearson (1991), Poellner (2009), Gemes (2009), and Richardson (2009) on the point, but he might have added Ridley (2007, 2009), Owen (2002; 2007, 96-102), Janaway (2009), Pippin (2009, 2010), and Clark and Dudrick (2009), and the later Katsafanas (2013) and Anderson (2013b) belong in the same line. These interpretations differ widely in the details (for discussion, see Anderson 2013b), but a broadly common core strategy is to separate a libertarian conception of free will that grounds moral responsibility, which is supposed to be rejected by Nietzsche, from some compatibilist-style idea of freedom as *control* in action, which he is supposed to accept as possible and even valuable.

8. Leiter has a curiously restrictive conception of what the philosophically interesting concept of freedom could be: the structure of his argument seems to assume that only a notion of freedom explicitly at odds with fatalism (via presupposing the principle of alternate possibilities) could be the one at stake in philosophical discussion (see, e.g., Leiter 2011, 115; also nn. 9 and 10). This leads to curious slippages in Leiter's argumentation, noted in section 3, below (see note 24 and the paragraph to which it is appended).

9. Leiter borrows the notion of "persuasive definition" from Stevenson (1938); the idea is that the redefinition changes the conceptual content attached to a word without altering its emotional and evaluative connotations, so as to exploit the old evaluative power on behalf of the newly defined idea.

10. Leiter (2011, 110-18) minimizes the number of such texts, but he is wrong. A partial list of texts in which Nietzsche endorses the value and/or possibility of freedom, autonomy, or

independence in some sense—sometimes making explicit that the relevant form of freedom needs to be compatible with necessitarianism—would have to include, in addition to the sections of *GM* that are our primary focus, *UM* IV, 11; *HH* Pref.; *GS* 98, 117, 335, 347; *Z* I, 17, 21; *BGE* 29, 41, 44, 46, 188, 201, 203, 212, 213, 284; *GM* II, 17 and III, 10; *TI* IX, 38, 41, and 49; and *A* 54. In many of these texts, the notion of autonomy is connected to the individual's ability to take responsibility for himself, which we will see to be central to the account in *GM* II, 1-3.

11. Williams (1993, 66) simplifies the basic conception of voluntariness down to the act's being intentional and done "in a normal state of mind," but the literature on what makes one's state of mind "normal" in the relevant sense is enormous, and I hasten to add that the list of possible defects offered in the text is radically incomplete. It is meant to gesture at a large domain of complications that have been explored in the literature about what it takes to sustain the the attribution of the act to "the agent herself," the judgment that she was responsible, claims that she was really "in" her action, and the like. For purposes of this paper, we can safely abstract from the details, though it is important for our purposes *that there are* subtle complications in the area. Do note that this characterization of the voluntary *does not* require deliberation as a necessary condition of voluntariness; it is required only that the action be connected in the right way to whatever deliberation there was.

12. Williams (1995 [1989], 27, and note 13) mentions Frankfurt and Taylor as writers with a broadly similar program, but there are many others, as I indicate below.

13. Irwin (1980, 144) adduces this very ambition as further support for his attribution of this complex account of responsibility to Aristotle, since it rationalizes the starting point of the *Ethics*

in a conception of happiness construed as an ultimate good providing a comprehensive order for practical life.

14. The list is meant to allude to versions of a similar strategy that have been (or might be) developed based on the work of Michael Bratman, Gary Watson, Harry Frankfurt, David Velleman, and Christine Korsgaard. Obviously, these versions would have quite different basic foundations, highlighting the remarkable flexibility of the basic strategy.

15. See Craig (1990) for an developed example of the approach. I am indebted to Elijah Millgram (pers. comm.) for calling my attention to this strand in Williams' own philosophical procedure.

16. Williams' point here builds on the deep idea behind the notion of "agent regret" he worked out in his thinking about moral luck (see Williams 1981 [1976]a, 27-33, *et passim*). As he powerfully shows, we often ought to take responsibility for things that happen in our lives through luck, even though the role of luck entails that the agent is not rightly held responsible (in the standard sense of moral responsibility). In such cases, it is the pull of the ideal of mature agency that issues the demand to take responsibility and induces the feelings of agent regret. Williams' famous example is of a lorry driver who hits and kills a small child who ran out into the road just as he was passing. As Williams argues, the gap between the demands of social/ moral responsibility and the more personal form of taking responsibility that gives rise to agent regret in such cases is precisely exposed by the role of insurance coverage. The insurance cover can completely discharge the sort of responsibility proper to anonymous social relations within a public, making the child's family whole for their loss in an abstract *legal* sense. But an excess

feeling remains behind—a need for some further gesture of responsibility that could never be discharged by the insurance company, but only by the driver herself. If she cannot bring herself to take responsibility by owning and expressing agent regret about the accident, we think less of her as a mature agent, even if there was no blamable negligence involved and all insurable debts have been discharged.

17. Nietzsche's genealogy of moral guilt in the Second Treatise is immensely complex and not as well understood in the literature as it should be. Unfortunately, space precludes even a sketch of the complex structure of the argument here, but readers can find a number of excellent pointers in Reginster (2011). I do not agree with all the details presented there (e.g., I doubt that a mature notion of personal worth is as present in *GM* II, 8 as Reginster suggests (at p. 69), and likewise that the lawbreakers depicted in *GM* II, 9 are actually far enough down the path toward guilt, psychologically speaking, to accept their punishment in any kind of restorative spirit, as suggested at p. 74). Still, the paper does a nice job of recognizing how complex Nietzsche's argument is, and identifying many of the key themes that need to be put together in order to arrive at an adequate reconstruction. Of particular importance for our purposes is Reginster's clear recognition that the moral psychological achievement attained by the sovereign individual when he develops a *conscience* needs to be remobilized in an essential but extremely delicate role within Nietzsche's account of the "moralization" (*GM* II, 21) that transforms the feeling of debt/guilt into the highly destructive form he aims to criticize.

18. Notice, too, that the quoted text presented it as entirely non-obvious that the sovereign individual would even need a word for this conscience. If promising were a key locus of

(potentially conflictual) interpersonal relations of holding one another responsible, then such a word would *clearly* be needed to facilitate negotiations and call out the salient obligations. We can infer that it was no accident or oversight on Nietzsche's part to leave out the others who would hold the individual responsible; he does not need them because his story is not about holding responsible at all, but something else.

19. Hatab (2008, 78) rests a good part of this case on reading the treatment of the sovereign individual in light of another history of morality outlined in *BGE* 32; Acampora (2013, 135-6) also appears to endose this reading. That text separates "a period that one may call *moral* ["*moralische*"] in the narrower sense" from distinct "*pre-moral* ["*vormoralishe*"]" and "*extra-moral* ["*aussermoralische*"]" stages. The three-way distinction has to do with the basis for evaluating actions: in the early stage, the value of an action resides entirely in its consequences; in the properly moral stage, the value is located in the *origin* of the action, which eventually comes to be identified in a privileged way as the *intention* of the agent, and the looked for extra-moral phase is supposed to evaluate actions in terms of what is *unintentional* in them, treating actions as symptoms of the basic condition of the person. Hatab is right to locate the sovereign individual at the "moral" stage in this alternative historical classification, since his concern is with the *origin* of his actions and expressions in the self for which he takes responsibility. But that result does not yet locate the sovereign individual within modern altruistic morality, for that latter structure is only *one special class* of the morality focused on the origin of actions—namely the phase that restricts the relevant origin to *intentions*, together with problems of praise and blame and holding responsible. In *BGE* 32, that version is separated from a more "aristocratic"

version that treats the question of origin as a matter of "'descent'" (perhaps tracing the value of action to an origin in the noble or base character of the person who did it). As I show in the text below, the sovereign individual would belong there, not in the morality focused on intentions and blame, and so he cannot be identified with the "modern ideal of individual rational autonomy" (Hatab 2008, 76).

I note that Nietzsche never mentions the morality of custom (or uses the word '*Sittlichkeit*') in *BGE* 32, and the classification offered there seems to me to cut across the distinction between the morality of custom and its successor: a post-morality-of-custom system could concern itself exclusively with consequences in the "*vormoralische*" mode (*sensu BGE* 32), and a morality of custom could attend to the source of actions in the agent, as suggested by *Genealogy*'s mention of "the oldest and most naïve moral canon ["*Moralkanon*"] of *justice*" (*GM* II, 8), namely that "everything [every action] has a price," which seems capable of operating within either *Sittlichkeit* or *Moralität*.

20. Here is the most relevant sentence:

> The "free" human being, the possessor of a long, unbreakable will, has in this possession his *standard of value* as well: looking out from himself toward the others, he honors or holds in contempt; and just as necessarily as he honors the ones like him, the strong and reliable (those who are *permitted* to promise),—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 even 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 already the moment it leaves his mouth. [*GM* II, 2]

As ever, there is much one could say about this long sentence, but on the immediate point, the implication is crystal clear. Whatever else it is, this deployment of a "*standard of value*" is clearly valuation "out of the *pathos of distance*" (*GM* I, 2), which is the defining mark of noble valuation in the First Treatise sense. There is one set of expectations and standard of values that is apt for those who are permitted to promise, and a different set for those who are not; the basic egalitarian/universalist application of valuation that defines the good/evil pattern of values is thereby violated (or perhaps better, rejected).

21. Acampora emphasizes the thoroughly naturalistic (strictly causal, non-normative) and deflationary character of Nietzsche's account of the origins of the power of memory as part of her case that the capacity to promise should be understood in a similar way. To my eye, however, promising is treated as a further achievement.

22. Nietzsche emphasizes the naturalistic ambitions of this background account when he opens the Second Treatise with the question: "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).

23. It might appear from the quoted passage that Leiter's epiphenomenalism is restricted to the will specifically, but the paper of his own cited within the quotation (Leiter 2007) suggests that the view is intended to extend to a global epiphenomenalism about conscious states.

24. [[[ADD ex. re door locking promise]]] Ironically, Leiter's own theory of Nietzsche's supposed practice of "persuasive definition" itself apparently depends on non-epiphenomenal conscious causation with practical effects running through "the will." The idea, after all, was that Nietzsche exploits the positive connotations of terms like 'freedom' and 'autonomy' in order to "activate the causal levers of at least some readers that will lead them toward this new ideal" (Leiter 2011, 112). But no matter how opaque they are to themselves about the emotional resonance of the target hot-button terms, if these readers do not understand, consciously, what the "new ideal" is supposed to be, and do so based on their comprehension of Nietzsche's text, and then attach themselves to the ideal practically because of that comprehension, then what happened to them does not count as persuasive definition.

25. See Dannenberg (2015), discussed below, to which I am indebted in this part of the discussion.

26. Nietzsche's explanation of the importance of forgetting also cuts directly against Leiter's epiphenomenalism regarding the causal role of conscious states in the production of action. Nietzsche explains that "Forgetting is no mere *vis inertiae* as the superficial believe; rather it is an ative and in the strictest sense positive faculty of suppression." Crucially, the reason this faculty is so important for us is that it provides "a little *tabula rasa* of *consciousness* [my ital.] so that there is again space for new things, above all for the nobler functions and functionaries, for *ruling* [my ital.], foreseeing, predetermining (for our organism is set up oligarchically)—that is the use of this active forgetfulness, a doorkeeper as it were, an upholder of psychic order" (*GM* II, 1). On this story, the function of the active faculty of forgetting is precisely to create the

space in which *conscious* states are able to *rule* the psychic oligarchy so as to plan and control its future actions, rather than being pushed around by other drives and attitudes forcing their way into consciousness, or determining how things go in general. Even though a few texts hint at the idea, strict epiphenomenalism seems not to have been under serious consideration by Nietzsche as a candidate stable view (on the will in particular, see also *BGE* 203).

27. The apparent idea behind Nietzsche's contrast is that those other "things and circumstances" could be represented by a different (cognitive) faculty, while the will remained focused on its original intention without interference.

28. Nietzsche frames the inquiry into memory in a way that indicates its merely instrumental role in the larger story. He introduces the discussion of ordinary memory with a question: "'How does one make a memory for the human animal?'" But that question (and hence the entire discussion) is located within a wider instrumental context by its setup: Nietzsche first notes that a notion like conscience "has behind it a long history and metamorphosis" and that the sovereign individual is "a ripe fruit, but also a *late* fruit"; he then remarks,

> how long this fruit had to hang on the tree harsh and sour! And for a still longer time one could see nothing of such a fruit,—no one could have promised it, however certainly everything on the tree was prepared for it and in the process of growing toward it!—"How does one make a memory for the human animal?" … an age old problem…

The clear suggestion is that the making of memory belongs among the "unripe," "sour," "preparations" in the "long history" leading up to the achievement of the sovereign individual's conscience. (All quotations from *GM* II, 3.)

29. This is among the points being made in Williams' (1981 [1976]b) famous example about saving one's wife in preference to an equally endangered stranger—where the context is the "one thought too many" objection to moral theory's handling of special relationships and the impartial demands of morality. My wife has a right to expect of me that I will save her without thinking twice (or even once) about it *in that* she has a right to expect such immediate, unreflective evaluative reactions from me (e.g., that her danger will immediately prompt my effort to help, that my concern for her will immediately override impartial concern for all in emergency situations, etc.).

30. See Dannenberg (2015, 160-2). The key sources of skepticism about such promises are 1) that it is hard to understand how I am genuinely *bound* by them (and in striking contrast to ordinary promises to others, it seems acceptable and even normal to reconsider the wisdom of promises to oneself in light of new information), and 2) that the very idea of promising myself seems to suggest a troubling lack of trust or confidence in myself.

31. [SHARPEN CONTRST; thx BR] See Ridley (2009) for an alternative account of what is special about what the sovereign individual promises, which focuses on a contrast between "the spirit" of the promise and its "letter." While there are some similarities between our claims, he connects the matter to his version of an expressivist theory of action, and I prefer to avoid that commitment. [[[Add K.]]] what is called for when circumstances prevent our performing on promises to a true friend is not compensation and guilt, but Williams-style agent regret and a forward-looking reaffirmation of the friendship's value to us; we do *judge* people who cannot bring themselves to express such agent regret or repair their close relationships against the

standard of mature agency, but we treat the needed repair as a matter for the friends themselves, and do not—or anyway, *should* not—see it as a matter for a court, or impersonal communal moral sanction

32. Owen (2007, 101-2) offers a treatment of this point developing Ridley's ideas.

33. The idea is that my behavior is autonomous when it is active, not passive, and it is active when the causal links connecting me into the necessitarian nexus run "inside-outward," making me a causal agent rather than a patient. The paradigm example is reasoning itself, where transition from one state to another is governed by the law of reason expressing my essential power, but we might also think of ourselves as approximating causal agency more, the better we understand the world in general, since our understanding thereby effects a kind of identification with the causes of things and their intelligible order.

34. A paradigmatic case is the discussion of certain ascetic instincts of the philosopher at *GM* III, 8, but there are many examples. Other relevant texts include *WP* 46 (*KSA* 13: 394, 14[219]), 778; *CW* 8; and *BGE* 6, 231.

35. Views attributing any kind of essence to individual selves or human "types" also run afoul of a very large line of Nietzschean texts rejecting the very idea that anything at all ever has a metaphysical essence fixing its identity conditions. For a comprehensive and convincing assembly and discussion of this strand in Nietzsche's thinking, see Poellner (1995, 79-111).

36. For a concise and persuasive exploration of reasons against understanding Nietzsche's conception of the unity of the self in terms of the rule of a dominating drive, see Katsafanas (2011, 98-100).

37. One locus for this theme that is highly salient for our purposes arises in Nietzsche's exploration of the psychological effects of socialization at *GM* II, 16.

38. Richardson (1996, 44-52, *et passim*) offers a compelling account of what is involved in the incorporation of one drive by another, according to Nietzsche.

39. Pers. comm.

40. Nietzsche's pervasive non-rationalizing explanations of behavior in terms of drives make this problem highly salient. Very often, such explanations make the activity seem alienated rather than autonomous, as for example, when a compulsive eating drive induces me to get up from writing and head to the kitchen in search of leftover dessert despite my being full and having resolved to eat less. What justification can we offer for treating the eating drive as "external" rather than "internal" in such a case? Appeal to the "dominant drive" seems singularly unhelpful, since *at the moment, at least,* the eating drive itself appears to be dominant (compare *GM* III, 8). If we are to deny it that designation (despite its effectiveness), we will have to do so by appeal to an identification of some other drive structure with the "real" self, by reference to which the eating is akratic. But what justifies that identification, once we have given up on the Spinozistic identification of the true self with the intellect and taken on board Nietzsche's skepticism about any fixed, given ego and any metaphysical essences?

41. Ridley (2007) also defends this thesis.

42. The discussion in the next five paragraphs is adapted from my earlier treatment in Anderson (2013b).

43. A similar idea was suggested already by Nehamas (1985, 186-7).

44. Nietzsche makes this account explicit in a passage from the notebooks that also indicates his skepticism about any traditional theory of the will:

> *Weakness of the will*: that is a metaphor that can prove misleading. For there is no will, and consequently neither a strong nor a weak will. The multitude and disgregation 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 the lack of gravity; in the latter, the precision and the clarity of the direction. [*WP* 46 = *KSA* 13: 394, 14[219]]

See also,

> Fear of the senses, of the desires, of the passions, when it goes so far as to counsel us against them, is already a symptom of weakness: extreme measures always indicate abnormal conditions. What is lacking, or crumbling, here, is the strength to restrain an impulse: if one's instinct is to have to succumb, i.e., to *have* to react, then one does well to avoid the opportunities ("seductions") for it. [*WP* 778 = *KSA* 13: 341, 14[157]]

45. I defend this interpretation through a detailed interpretation of *BGE* 12 in Anderson (2012, 211-16).

46. Nietzsche himself brings together his integration-based conception of strength and weakness with the thesis of psychological anti-atomism and the problem of boundaries for the self in a remarkable passage from the notebooks:

> *No* subject "atoms." The sphere of a subject constantly growing or decreasing, the center of the system constantly *shifting*; in cases where it cannot organize the appropriate mass, it

> breaks into two parts. On the other hand, it can transform a weaker subject into its functionary without destroying it, and to a certain degree form a new unity with it. No "substance," rather something that as such strives after greater strength, and that wants to "preserve" itself only indirectly (it wants to *surpass* itself—). [*WP* 488 = *KSA* 12: 391-2, 9[98]]

Here, strength and weakness are implicitly understood in terms of the degree of integration among the constituents of the psychological entity, which is not an "atom" or "substance," but instead a shifting "center," whose boundaries ("sphere") vary both with internal changes in the degree of unification and with efforts to incorporate "weaker subjects" from the environment. The possibility of "fission" indicates both the vulnerability of such "centers" to weakness, and the fact that their identity is constituted by their strength/unity, as I suggested in the text. A relative center of strength is thereby able to mark out a psychological unit for analysis, even though its boundaries (and consequently also the exact "center") are "constantly shifting." In this way, the very radicalness of the anti-atomism helps Nietzsche to address the problem of the internal/external boundary.

47. See Anderson (2012) for a defense of this thesis.

48. Nietzsche's own ideas about the problem of selfhood are bound up with some of his most speculative theorizing about the basic elements of psychological life, but in its most abstract form, the structure of his theorizing—traceable to the ideas of integrated unity and the internal/external distinction—seems to me worth some investment. Recent efforts by other philosophers

(e.g., Michael Bratman, 2007) to identify the voice speaking for "the agent herself" with the one that can unify her life suggest both the fruitful variations and the promise of the general strategy.

49. See Pippin (2010, 78-9). Elsewhere (Anderson 2013a; 2013b, 452-4), I have raised some doubts about the wider expressivist theory of action within which Pippin locates this point, but on the hermeneutic thesis at stake here, we are in entire agreement.