9

What is a Nietzschean Self?[1]

R. Lanier Anderson

1. INTRODUCTION: NIETZSCHE AND KANTIAN ETHICS

I am among those who see the history of nineteenth-century philosophy largely as a story about Kantianism. German thought of the period was dominated by strands occupied with working out the challenges Kant posed and exploring the resources of his system. Moreover, these lines of thought put German philosophy for the first time really on the map, indeed arguably at the centre of the map, of the European intellectual world. The point is perhaps clearest in theoretical philosophy, where even avowed positivists at least frame their programs by reference to Kantian problematics—witness Ernst Mach's anecdote about being pulled into philosophy by the *Prolegomena* or Richard Avenarius's use of the title '*Kritik der reinen Erfahrung*'.[2] The parallel point on practical philosophy's side of the street is perhaps more controversial. Kant's shadow can seem short if one focuses on the emergence of utilitarianism as a strong competitor to Kantian theory, and even to the underlying deontological intuitions at its basis. But on the side of Kantian influence, one can cite—well, first of all, the fact that Kant's distinction between theoretical and practical philosophy substantially informs this very way of distinguishing the two sides of the street—but also the academic spread of German idealism and its

[1] This paper was written for Christopher Janaway's workshop on 'Nietzsche and Kantian Ethics' at the University of Southampton (April 2009). It benefited greatly from the discussion and from written comments of the workshop participants, to whom I am grateful. I also received helpful audience feedback from the Aesthetics Workshop and the William James work-in-progress group (both at Stanford), the Philosophy Colloquium at the University of Illinois, and the audience at the University of New Mexico graduate philosophy conference. I benefited from especially detailed written comments by Elijah Millgram, and by my colleague Allen Wood, who saved me from some errors, but who will doubtless find that many more survived into this version. Finally, the paper was improved through conversations with Will Beals, Chad Carmichael, Ken Gemes, David Hills, Chris Janaway, Paul Katsafanas, Joshua Landy, Elijah Millgram, Alexander Nehamas, Peter Railton, Simon Robertson, Richard Schacht, Tamar Schapiro, Ben Wolfson, and Katherine Preston.

[2] For Mach's statement about Kant's influence on his development, see the long note at Mach (1910 [1886]: 23–4). See also Avenarius (1888, xi–xiii, *et passim*).

broader reception in romanticism, the influence of its 'left Hegelian' and other early critics, the mid-century popularity of Schopenhauer's pessimism, the variegated 'back to Kant' movement, and the gradual emergence from Kantian roots of something very like our own philosophical problem about the place of normativity in a naturalistic worldview.[3]

Nietzsche's place in this landscape remains seriously contested. At first glance, it might seem strange that there is any occasion for debate here. Nietzsche obviously rejects core principles of Kant's moral theory, particularly its account of the categorical imperative and the moral argument for God, freedom, and immortality as postulates of practical reason (see *GS* 335; *BGE* 5, 11, 187; *GM* II, 6, and III, 12, 25; *A* 11; *et passim*).[4] In this case, moreover, Nietzsche's criticisms are not restricted to mere name-calling or hyperbole, but cut to the core of his philosophical concerns. Whereas Kant accepts at face value the normative force of ordinary moral intuition—and indeed, takes it as a sufficient basis for a regressive argument to establish the fundamental principle of morality—Nietzsche, by contrast, offers a debunking genealogy of the same intuitions, designed to expose our attachment to them as so much (unattractively) motivated believing. In addition, Nietzsche raises sceptical objections against the underlying moral psychology needed to make sense of Kantian moral theory, attacking notions like the will, pure practical reason, the alienating effects of inclination, etc.[5] Thus it is important to acknowledge Nietzsche's anti-Kantian sensibilities from the outset.

All that said, there remain noteworthy parallels between key ideas of the Kantian tradition in ethical thought and apparently fundamental commitments of Nietzsche's. Perhaps the most striking point of contact concerns the value of autonomy.[6] Autonomy, for many Kantians, is not only important as an idea of our freedom, but also serves as an *ideal*. Autonomous agency itself is what carries value beyond any price, demands respect in our dealings with others, and

[3] For discussion of the Kantian roots of discussions of the problem of naturalism and normativity, see Hatfield (1991) and, for neo-Kantianism specifically, Anderson (2005b).

[4] Citations to Nietzsche's texts will be made parenthetically, using standard abbreviations as noted in the references; I have made use of the translations and editions detailed there. I also cite Kant's *Critique of Pure Reason* using the standard A/B format to refer to the pages of the first (=A) and second (=B) editions.

[5] As Williams (2006 [1993]) points out, these two forms of criticism are deeply connected. Nietzsche defends a minimalist moral psychology by first identifying the respects in which the apparently implied psychology of ordinary moral intuitions makes commitments in 'excess' of what a cold-eyed 'realistic' apprehension of human behaviour in other domains would require (Williams 2006 [1993]: 302); he then deploys a genealogical 'hermeneutics of suspicion' to undermine confidence in the moral intuitions. The argument thereby suggests that the psychological commitments are rationalizing fantasy, rather than necessary postulates of reason. For further discussion, see Section 3.

[6] For discussion of the relation between normativity and autonomy in nineteenth-century thought and a particularly intriguing discussion of Nietzsche's conception of autonomy, see Reginster (2012). At a more abstract level, Hill (2003: 196–229) argues for Kant's influence on Nietzsche's conception of the general structure of the problem space for moral theory, in addition to more substantive parallels like the agreement about the value of autonomy mentioned here.

so on. It is both the source of morality's authority over our actions and the basic value that morality strives to protect. While Nietzsche denies that autonomy comes built in as standard equipment along with humanity or rationality as such, he nevertheless seems to share the Kantian (or anyway, post-Kantian) emphasis on its value. Autonomy is central to the rare form of strong individuality he praises: the free spirit he idealizes is supposed to be independent from custom and tradition; she 'creates herself' precisely in the sense of giving herself values or laws of her own; she has 'independence of the soul' (see *GS* 98, 99, 335, 347; *BGE* 29, 41, 43–4, 203; *GM* II, 1–3; *et passim*). Of course, Nietzsche does not conceive autonomy along orthodox Kantian lines, and he consequently rejects the Kantian claim that recognizing the value of autonomy by itself constrains us to accept the full content of altruistic morality. But just here, even Nietzschean immoralism can be understood as indebted to the post-Kantian tradition. After all, Nietzsche's complaint sounds a note remarkably similar to the famous Hegelian objection that Kant's moral theory is a 'mere formalism', lacking sufficient content to entail the substantive demands of morality. In fact, I have always thought that the 'mere formalism' objection offers a surprisingly illuminating way to sketch one key aspect of Nietzsche's normative stance, along lines like this: Kant successfully identified what should have basic value for us, namely autonomy, but Hegel was right that such a 'merely formal' value cannot possibly entail all of traditional altruistic morality, and (now contra Hegel) that is a good thing too, since the 'un-selfing' tendencies of such morality are fundamentally bad for us. What Hegel saw as a bug is actually a *feature*, indeed the real and deep insight, of Kantian moral theory.

With this, we come face to face with a difficulty. How can we reconstruct the philosophical shape of a value theory that seems at once fundamentally anti-Kantian but also built on a core of broadly Kantian ideas? I will explore one way this dilemma plays out in certain details of moral psychology.

2. NIETZSCHEAN MORAL PSYCHOLOGY: NATURALISM VERSUS TRANSCENDENTALISM

Kantian ethical theory in general, and its conception of autonomy in particular, rests on a crucial assumption about moral psychology. For Kantians, there is a fundamental difference between two types of motivational incentives—those of reason and those of inclination—and in any context of action or decision, reason is supposed to have the basic capacity to 'stand back' from the biddings of inclination and decide independently whether the inclination is to be endorsed by the self or not.[7] Our autonomy depends on this capacity to 'stand back' from

[7] Of course, Kant need not, and does not, deny that reason and inclination may interact in the same attitude, e.g. to form passions in which our inclinations are informed by influence from our power of choice. The key point for my purposes is just that the separation of two sources of

desires and assess them, so that the self can follow reason's law even when it stands completely athwart the demands of all our inclinations. But just here we might worry, from a Nietzschean point of view, since it is a matter of controversy *whether there is even any such thing*, for Nietzsche, as a self capable of 'standing back' from our conative attitudes in this fashion.

Naturalist readers of Nietzsche such as Brian Leiter (2002, 2009; also Leiter and Knobe 2007) and Matthias Risse (2007) insist that there is not.[8] They emphasize the many texts that express Nietzsche's sceptical, or perhaps even eliminativist, position about the self.[9] According to this strand of thought, our belief in a unified conscious self over and above our desires, drives, or inclinations is an illusion. In fact, the self is nothing but a 'social structure of the drives and affects' (*BGE* 12), and we 'deceive ourselves' about this multiplicity when we take it as a unified, substantial thing 'by means of the synthetic concept "I"' (*BGE* 19). When it appears to us that our conscious self or intellect has taken some basic decision against a drive or other conative attitude within us, in reality what occurs is merely that *another drive*, which is opposed to the first and, more dominant, has seized the place of speaking for the self (*D* 109). While we often suppose that the intellect is 'something that is essentially opposed to the instincts', in fact (contrary to the Kantian assumption) 'it is actually nothing but a *certain behavior of the instincts toward one another*' (*GS* 333). On this picture, so far from there being a self capable of standing back from all the drives, what speaks for 'the self' is nothing but the strongest or dominant drive itself.[10]

motivation allows Kantians to claim that it is always (motivationally) possible for an agent to 'stand back' from inclinations altogether, assess them from the standpoint of reason alone, and act in a way that is motivated by pure reason. See Reginster (2012) for additional discussion. (Thanks to Allen Wood for clarifying exchanges.)

[8] Of course, many of Nietzsche's French post-structuralist readers (e.g. Foucault, Derrida) are equally keen to emphasize scepticism about any substantial notion of the self. Given the notable differences between the naturalist and post-structuralist camps in background philosophical motivations, it is remarkable in its own way that they share such a prominent investment in this interpretation of Nietzschean doctrine.

[9] Just to provide a hint of the domain, here is a quick and dirty, radically incomplete selection of Nietzsche's comments in this vein: '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 a 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).

[10] This last interpretive position—that the Nietzschean self is just the strongest drive—is widely endorsed by commentators even outside the naturalist and post-structuralist camps; see e.g. Reginster (2003).

By contrast, Kantian-inspired readers such as Sebastian Gardner (2009) insist that, notwithstanding the sceptical strand of texts just canvassed, Nietzsche's own practical philosophy *commits* him to a conscious self capable of 'standing back' from the drives in a broadly Kantian sense. For Gardner, Nietzsche's thought contains 'a buried transcendental dimension' (Gardner 2009: 19) with substantial implications for moral psychology. Consider, for example, the preconditions for the 'creation of values' central to Nietzsche's value theory. In order for the individual to create values of her *own*, the thought goes, she must have a conception of herself as a *unified* practical agent who is the source of those values. Even if the values she posits are influenced by the drives within her, the individual self must (first-personally) think of them as her own—and not merely the demands of some dominating drive—on pain of a 'profound self-alienation' (Gardner 2009: 9) which would undermine the very autonomy Nietzsche sought to secure by appealing to the creation of values in the first place. I confess that this argument strikes me as potentially question-begging against the Leiter-style naturalist. (It seems that the naturalist can simply deflate the autonomy Nietzsche sought along with the notion of selfhood, insisting that when 'I' speak the values of the dominant drive in the voice of my (more or less illusory) self, that is all the autonomy, and all the 'first-personalism', that Nietzsche wants or needs.) In any case, the result seems to be based primarily on an a priori argument identifying alleged presuppositions of Nietzschean positions, rather than any direct argument from Nietzsche's texts.[11] As such, it might be thought to tell us more about the shape and force of *Gardner's* post-Kantian commitments than it does about Nietzsche's own view.[12]

[11] The same basic form of argument, which posits an autonomous self as a precondition of practical agency quite generally, is a widespread move in the Kantian tradition. For a classic example, see the well-known response to Parfit in Korsgaard (1996: 363–97).

[12] I should note, in addition, a second kind of argument for the transcendentalist conclusion in Gardner, to which I have a similar reaction. The second argument focuses on whether a mere collection of drives could even generate the requisite *idea* of a unified 'I' without actually *being* a unified transcendental self of the sort in dispute. Gardner writes 'So the question arises, how, except in the perspective of an I, of something that takes itself to have unity of the self's sort, can a conception of unity sufficient to account for the fiction of the I be formed? (As it might be put: How can the 'idea' of the I *occur* to a unit of will to power or composite thereof—or to anything *less* than an I?)' (Gardner 2009: 69). I am puzzled by Gardner's puzzlement here. Three ideas suggest themselves. Perhaps, first, the worry is just a version of the problem of (the unity of) consciousness—that is, a doubt that the collection of subpersonal attitudes Nietzsche postulates in the self could ever give rise to any (unified) conscious state at all. But this worry has nothing special to do with the representation 'I'; it would apply in the same way to any representational content accompanied by reflective consciousness. Since Nietzsche seems to be willing to assume fairly substantial representational capacities for his drives and affects, he is perhaps better positioned with respect to this general problem than other radically naturalistic positions. If, second, there is supposed to be a *specific* problem about a collection-self coming up with a particular *content* of representation, the 'I', then I confess that the argument strikes me as being parallel to Descartes' Med. III proof of God's existence, and subject to similar problems. Some representation (<God>, <I>) is supposed to be so special that a representational system could not reach it by any kind of extrapolation or invention, so we must conclude that the object of the representation really exists,

But a similar point is raised by Chris Janaway (2009) in an explicitly text-based context that *does* tie the result to distinctive Nietzschean doctrine and not just general Kantian principle. As Janaway insists, Nietzsche's perspectivist conception of objectivity requires the cognitive self to 'stand back' from its affects in much the sense under discussion. Objectivity is to be seen:

> not as 'contemplation without interest' (which is a nonsensical absurdity), but as the ability to *control* one's Pro and Con and to dispose of them, so that one knows how to employ a *variety* of perspectives and affective interpretations in the service of knowledge... [T]he *more* affects we allow to speak about one thing, the *more* eyes, different eyes, we can use to observe one thing, the more complete will our 'concept' of the thing, our 'objectivity,' be (*GM* III, 12).

Here the cognitive self that does the 'controlling' cannot plausibly be identified with some dominant affect or drive. For if the self were just the dominant affect, then *that* affect, at least, would not be 'controlled' and 'disposed of' by an independent cognitive self, and the wanted objectivity would not be achieved. Perspectivist objectivity thus apparently requires a capacity on the part of the cognitive self to detach itself from its constituent drives and affects so as to take up attitudes towards them—even to control and manipulate them. Arguably, similar implications attach to other central Nietzschean ideas such as his ubiquitous emphasis on self-mastery and self-overcoming, or the 'sovereign individual' praised at *GM* II, 2.[13]

and has provided the representation's content through being perceived (or in some other way?). But what is so special, really? Supposing that the Nietzschean bundle-self could represent at all, why couldn't it manufacture for itself an illusory 'synthetic concept "I"' (*BGE* 19), and (falsely) think of itself under that concept? Perhaps, third, there is supposed to be a deep Kantian reason that all representation (or at least reflective representation) necessarily *presupposes* a transcendental ego. For example, a Kantian might insist that representations can only come together and count as a *judgement* by being synthesized, and thereby brought into a unity through the activity of a single, conscious cognitive agency. This point, however, strikes me as more *Gardnerian/(post-)Kantian* than *Nietzschean* in flavour. That is, if some such thing is true, why should we receive the point as an interpretation of Nietzsche, rather than a *criticism* that he has overlooked a deep and important insight of transcendental philosophy? (Thanks to Christine Lopes and Allen Wood for clarifying exchanges on this last point.)

[13] Nietzsche's discussion at *GM* II, 1–3 provides fairly decisive evidence for the point, it seems to me. For recall, the distinctive capacity of the 'sovereign individual', promising, abrogates the normal forgetfulness that characterizes the experience of others, and the individual does this precisely by means of an *act of will* that persists across arbitrary psychological changes in which other drives are activated, and thereby instantiates a form of active *self-control* that is not interrupted by those other drives: 'a promise... is thus by no means simply a passive no-longer-being-able-to-get-rid-of the impression once it has been inscribed, not simple indigestion from a once pledged word over which one cannot regain control, but rather an active no-longer-wanting-to-get-rid-of, a willing on and on of something one has once willed, a true *memory of the will*: so that a world of new strange things, circumstances, and even acts of will may be placed without reservation between the original "I want", "I will do", and the actual discharge of the will, its *act*, without this long chain of the will breaking' (*GM* II, 1). Thus, the sovereign individual is a possible type defined by the capacity of a *whole self* to assume a diachronically stable attitude of commitment, which persists through the alterations of the individual drives and controls action even in the face of their vicissitudes. That is,

Our understanding of Nietzsche's moral psychology thus faces a genuine dilemma. On the one hand it is impossible to ignore the texts expressing scepticism about any substantial notion of the self, and even suggesting a reduction of the self to subpersonal drives and affects. But on the other, core Nietzschean ideas like self-overcoming and perspectivist objectivity seem to require some notion of a self separate from the drives.

To lay some of my cards on the table, this paper aims to carve out a 'third way' between naturalist and transcendentalist readings. One idea in the background will be the thought—pushed already by Nehamas (1985) and Schacht (1983: 306–9), but recently developed by others, including Janaway (2009), Gemes (2006, 2009b), and myself in earlier work (Anderson 2006)—that the Nietzschean self is not simply given as standard metaphysical equipment in every human, but is rather some kind of *task* or *achievement.*[14] My strategy will be to surround this suggestion with enough moral psychological details to fill out a viable competitor to the naturalist and transcendentalist conceptions of selfhood that have received greater development in the philosophical tradition, e.g. from Hume, Kant, and their followers.

3. HOW MINIMALIST *IS* NIETZSCHE'S MORAL PSYCHOLOGY?

The tendency in the literature to identify the Nietzschean self with its strongest drive has become increasingly pronounced since Bernard Williams' enormously influential 1993 paper 'Nietzsche's minimalist moral psychology' (Williams 2006 [1993]).

Let me note immediately that Williams' paper itself took no firm position on the general nature of the Nietzschean self or its relation to the drives. His agenda was shaped *not* by the reductionist aim to identify the self with some subpersonal constituent(s), but rather by an Edward Craig-inspired program of reconfiguring central philosophical notions in light of connections to their genuine social function and the needs they fulfil (Craig 1990).[15] In line with that program, Williams emphasized Nietzsche's broad suspicion against the 'excess of moral content' (Williams 2006 [1993]: 302)—i.e. content beyond what is justified by their core function—carried by many moral psychological notions. In particular,

what characterizes the type is precisely that there is a difference between the self as a whole and the variable drives.

[14] Allen Wood (personal communication) rightly points out that this idea is not unique to Nietzsche, but also has a well developed life in the post-Kantian tradition going all the way back to Fichte.

[15] Thanks to Elijah Millgram for reinforcing to me the importance of this context, and to David Hills for discussion.

he focused on the *will*, construed as a simple faculty capable of causing results by prescription (i.e. simply by issuing an imperative that things should be so). Williams traces the 'moral excess' built into this notion to the way it is fine-tuned to match our need to assign moral blame. As he notes, 'Blame needs an occasion—an action—and a target—the person who did the action and who goes on from the action to meet the blame' (Williams 2006 [1993]: 307). The faculty of will nicely supplies these requirements, since its conceptual form includes an occasion (the willed action) and a subject/target, who caused the action by prescription, and who can therefore be assessed with blame to the exact degree that the outcome was in his power. In this sense, it is plausibly 'the needs, demands, and invitations of the morality system . . . [that] explain the peculiar psychology of the will' (Williams 2006 [1993]: 307), and that is enough to raise the suspicion that belief in the will arises not in response to general theoretical demands of psychological explanation, but instead from certain *desires* (or other pro-attitudes) rooted in 'the morality system'. If so, then it counts as a motivated belief, and deserves to be stripped out of a more realistic psychology.

Just here, though, more reductionist motivations can enter the picture, and in my view, such motivations have decisively shaped the paper's *reception* by Nietzsche scholars. Williams' approach clearly captures something important about Nietzsche's procedure, and it is natural to seek to generalize it—the will, after all, is only one example of the (allegedly) widespread effects of 'moralisation' (*GM* II, 7, 21) within commonsense psychology. The most tempting generalization strategy leaps from the rejection of the will, to the rejection of any specially posited power or faculty that seems to have a distinctive or important role in moral affairs, to the conclusion that 'minimalism' in this context should amount to something like modern-day 'Humeanism'—i.e. a general restriction of moral psychological explanations to a suitably austere ontological basis that permits appeal only to the (morally neutral) psychological attitudes of *belief* and *desire*. Such a strategy looks to have direct implications for the core capacity of the self at issue between transcendentalist and naturalist readings, namely the capacity to 'stand back' from our attitudes and endorse or reject them. Since that capacity fills a rather substantial moral role, it looks to be a reasonable target of Williams-inspired suspicion. Either it should be explained in terms of the minimal belief/desire apparatus or we should suspect that it, just like the will, is a fabrication of moral consciousness. The result lends substantial aid and comfort to reductionist or eliminativist readings of Nietzsche's sceptical remarks about any soul or self that would be separate from the drives: such readings answer to a minimalist demand by eliminating the self, or at least reducing it to the strongest conative attitude.

While tempting, a full-dress 'Humean' interpretation of what 'minimalism' requires cannot possibly be true to Nietzsche's intentions (nor, one last time, was that conclusion ever advanced by Williams). Compared to the ontologically stripped down, austere, well-nigh parched landscape of belief/desire psychology, Nietzsche's own moral psychological apparatus gives off a positively steamy air of tropical luxuriance. It is populated by an impressive array of attitude-types—drives, affects, instincts, desires, wills, feelings, moods, valuations, sensations, concepts, beliefs, convictions, fictions, imaginings, cognitions, and so on—and Nietzsche liberally appeals to the full range without evincing any noticeable concern about reducing apparently more complex attitudes (e.g. valuing) to simpler ones (e.g. desiring). In addition, in Nietzsche's actual usage, each attitude-type displays prodigious internal variety and complexity. For instance, Janaway (2009: 52) identifies at least thirty different affects playing explanatory roles in *BGE* alone, many of which are themselves identified in terms that appeal to further attitudes (e.g. the affect of *demanding* respect, the affect of the *feeling* of command). To consider another dimension, these attitudes can take very different kinds of objects as complements—from propositional contents, to individual objects, relations (e.g. 'rule over'), other attitudes, and even apparent abstracta (e.g. 'power'). Moreover, as I will argue below, the standard complement requirements for at least many of these attitude-types are themselves essentially more complex than those for ordinary beliefs and desires.[16] Finally, it is worth noting that Nietzsche himself takes the psychological reality constituted by these attitudes to be so nuanced and fine-grained as to outstrip (and by far!) the distinctions marked within his highly ramified explanatory apparatus—and indeed even all those available *in principle* to the capacity of conscious reflection (*GS* 335).

It appears, then, that the potential explanatory resources of Nietzsche's moral psychology are far greater than those we typically attribute to (or exploit within) a contemporary naturalist belief/desire psychology. Moreover, the added complexity to which Nietzsche helps himself seems entirely likely to survive the sort of minimalist program proposed by Williams. The postulated attitudes and their contents and objects are so luxuriantly complex precisely in the service of Nietzsche's efforts to capture the subtle variations of non-moral (and even *im*moral) psychological life. Thus they are highly unlikely to carry the sort of 'excess moral content' that Williams-style minimalism strives to remove.

We are now in a position to advance a more informative version of our problem: the question is whether Nietzsche's complex psychological apparatus provides the materials for a conception of the self that is separable from its constituent attitudes, in the sense of having the capacity to stand back from them

[16] We can at least begin to understand the kind of increase in complexity involved here by comparing it to the way some contemporary philosophers take valuing to be essentially a more complex attitude than desiring—valuing is often supposed to be some higher order attitude *built out of* and *referring to* desires, and therefore essentially more complex. See Michael Smith (1994: 130–47) for a helpful discussion of some of the options for understanding the relation between valuing and desiring in recent literature. Smith himself rejects the analysis of valuing in terms of desiring and argues (1994: 147–81) for an analysis resting on beliefs about normative reasons.

to assess them, endorse or reject them, 'control', and 'dispose of' them (*GM* III, 12) in the way that seems to be involved in the achievement of autonomy.

4. IS NIETZSCHEAN SCEPTICISM ABOUT THE SELF REDUCTIONIST? A READING OF *BGE* 12

I propose to make a preliminary assessment of Nietzschean scepticism about the self through a relatively close reading of one[17] of Nietzsche's most famous sketches of what a demystified conception of the 'soul', or self, might look like:

> As for materialistic atomism, it is one of the best refuted theories there are . . . thanks chiefly to the Dalmatian Boscovich . . . Boscovich has taught us to abjure belief in the last part of the earth that 'stood fast'—the belief in 'substance,' in 'matter,' in the earth-residuum and particle-atom: it is the greatest triumph over the senses . . . so far. One must, however, go further, and also declare war . . . against the 'atomistic need' which still leads a dangerous afterlife in places where no one suspects it . . . : one must also, first of all, give the finishing stroke to that other and more calamitous atomism which Christianity has taught best and longest, the *soul atomism*. Let it be permitted to designate by this expression the belief which regards the soul as something indestructible, eternal, indivisible, as a monad, as an *atomon*: this belief ought to be expelled from science! Between ourselves, it is not at all necessary to get rid of 'the soul' at the same time, and thus to renounce one of the most ancient and venerable hypotheses—as happens frequently to clumsy naturalists who can hardly touch on 'the soul' without immediately losing it. But the way is open for new versions and refinements of the soul-hypothesis; and such conceptions as 'mortal soul', and 'soul as subjective multiplicity', and 'soul as social structure of the drives and affects', want henceforth to have citizens' rights in science (*BGE* 12).

This passage is well known from its frequent starring role in support of naturalist readings that aim to reduce the Nietzschean self (in broadly Humean fashion) to a mere bundle of drives. On closer inspection, however, the text seems peculiarly miscast in that particular role. Four points are worth noting.

First, the official target of Nietzsche's attack is not the soul *per se*, but the *atomistic* theory of the soul, i.e. the view that the self is simple (i.e. without parts), and therefore indestructible or immortal. The argument from simplicity to

[17] For the purposes of this paper, I will focus on *BGE* 12 as a paradigmatic text, and I will not even attempt to interpret (or disarm) all of the textual evidence that Nietzsche held some eliminativist or reductionist view of the self. In fact, I think that many (though not all) of the texts, and very nearly all of the *published* texts, usually cited in support of such readings are quite a bit more equivocal than they seem to those who cite them. But treatment of the full range of textual evidence must await another occasion. Furthermore, I hasten to concede that at least some texts and notes in Nietzsche do suggest the sort of stronger reduction or elimination of the self that I fail to find in *BGE* 12, *D* 109, *BGE* 17 and 19, etc. My line on those texts will be that they are hyperbolic and do not reflect Nietzsche's considered position.

immortality goes at least all the way back to Plato, but it was a particular staple of early modern metaphysics, and that early modern version of the idea was the central result of rational psychology that Kant undermined in his 'paralogisms'. Nietzsche likewise rejects the conclusion of the traditional argument (hence his interest in the concept 'mortal soul' among the 'new refinements of the soul-hypothesis'). But what is more interesting for our purposes is Nietzsche's concomitant rejection of the argument's *premise.* The Kantian critique had already delegitimized the inference from the unity of consciousness to a simple, incorruptible, subjective *substance*, but Kant, followed here by the broad consensus of nineteenth century philosophical common sense, still insists on a very strong, logico-transcendental notion of the unity of consciousness, which, indeed, he takes to have far-reaching philosophical consequences (including *inter alia* blocking materialism in philosophy of mind).[18] By organizing his argument as an attack on *atomism* in psychology, Nietzsche clearly means to reject not just the inference to immortality, or to a *substance* underlying subjective consciousness, but also this strong notion of the unity of consciousness itself.[19] The main idea of

[18] Of course, Kant's position here must be qualified. In his view, the unity of consciousness does *not* permit any inference to the conclusion that the soul is a substance, nor that it persists beyond life (or outside the bounds of possible experience). That said, the 'merely logical' transcendental ego—i.e. the conception implicated in the 'I think' that plays such a key role in underwriting the unity and possibility of experience—is in fact simple and unified in a strong and consequential sense. In particular, its simplicity is part of the critical argument designed to cut off all materialism. Consider: 'Apperception is something real, and its simplicity lies in its very possibility. Now there is nothing real in space that is simple; for points (which constitute the only simple entities in space) are mere bounds, and not themselves something that serves to constitute space as parts. Thus, from this follows the impossibility of explaining how I am constituted as a merely thinking subject on the basis of *materialism*' (B 419–20). (Any spiritualist explanation is equally ruled out by critical strictures, of course; B 420.) The first edition 'Paralogisms' featured a much more indirect version of the view, but the argument still ultimately relies (albeit *very* indirectly, I admit) on the unity of consciousness, and consequent simplicity of the logical 'I think'; see A 383. But it was the more straightforward argument from the B edition that carried such enormous influence in nineteenth-century philosophy. (It is perhaps also worth noting, with a view towards note 20 below, that unlike Boscovich, Kant commits himself here to a continuum mechanical view of matter, which is constituted, not by points, but by the force exercised through space from points, and is therefore divisible/composite in principle in a way that makes it incompatible with the simplicity of apperception.)

[19] This result is the whole point of bringing up the opening discussion of physical science in the first place. Nietzsche's premise is that atomism qua doctrine has been definitively refuted in physical science (by insights of Boscovich, *et al.*), but that the underlying explanatory pattern persists, having spread to other theoretical domains such as psychology (the theory of the soul). He then argues from this premise to the conclusion that, in the absence of any support from analogy to a credible strategy of physical explanation, the overall atomistic way of thinking cannot claim to be driven by data or demanded by any principled a priori argument. On the contrary, it owes its plausibility solely to an '*atomistic need*', rooted perhaps in the thought-pattern's familiar similarity to our everyday representations of colliding stones, billiard balls, and the like. Now that the theoretical value of atomism as doctrine has been undermined in its core home area (physics), Nietzsche suggests, we should also 'go further' and reject its extension into psychology, which was always based more on atomism as need than on any substantial theoretical merits. The fact that the idea's main deployment in rational psychology was in the proof of immortality only increases the suspicion that it is so much motivated believing.

BGE 12 is to replace that notion with the hypothesis that even the basic self is essentially a complex ('soul as subjective multiplicity', 'soul as social structure of the drives and affects').

This result is quite problematic for any Gardner-style transcendental reading of Nietzsche's moral psychology, at least *as an interpretation of Nietzsche*. So far from accepting (what a Kantian would insist are) the transcendentalist implications of his commitments about human practical and cognitive capacities—as a 'buried transcendental dimension' of his thought (Gardner 2009: 19)—Nietzsche himself is determined to reject any such conception of the self. When he insists that the theoretical work of psychology could be done by a notion of the soul as a '*subjective multiplicity*' (*BGE* 12, my ital.), he means to deny, contra Kant, Gardner, *et al.*, that whatever is *subjective* at all must exhibit a strong and essential *unity* proper to consciousness as such, and thus to deny that there is any need to postulate a unified transcendental ego.

But second, the same thoroughgoing rejection of atomism from *BGE* 12 has striking implications that make trouble for a reductionist or eliminativist reading of his theory, as well. These implications concern the *relation* between the self and its drives and affects, given the sort of anti-atomism Nietzsche suggests. Consider, first, that as Richardson (1996: 44–52) points out, it is a basic feature of Nietzsche's theory of drives that they are capable of combining with one another to form larger, encompassing structures that count as drives in their own right, possessed of distinct aims and roles in the psychological economy, and thus some independence from their constituent sub-drives. (To take one of his examples, my drive for food and my drive to socialize can be integrated into a 'social eating' drive, which produces and governs its own distinctive pattern of behaviour (Richardson 1996: 47).) Note, secondly, that Nietzsche's rejection of the 'soul atomism' is meant to be conceptually parallel to a definite sort of criticism of materialistic atomism, which replaces indivisible atoms with a Boscovich/Kant system of point masses that fill space by exercising repulsive force through it. Following Lange, Nietzsche is relying on a dynamical, continuum mechanical interpretation of such a system, and it is *that interpretation* that has the radically anti-atomist implications. On this picture, matter consists essentially of attractive and repulsive forces operating from *points*; therefore it must be divisible 'all the way down',—division can simply reallocate the quantities of force (in which matter itself consists) along a continuum of geometrically available points.[20] Thus there are simply *no* material atoms. Now putting our

[20] In fact, Nietzsche's Lange-descended, continuum mechanical version of the view is not a good interpretation of Boscovich's actual theory. Boscovich does resolve matter into point-centres of force, but for him, matter itself consists in the *point-centres*, not the *forces* that operate from them. As a result, Boscovich *does not* in fact dispense with 'particles' in the sense Nietzsche intends. On the contrary, he explicitly treats these centre-of-force points as *indivisible* precisely because they are perfectly simple. Thus the ultimate constituents of matter (for Boscovich) are explicitly supposed to be just what Nietzsche says they are *not*, namely 'indestructible, eternal, indivisible' (*BGE* 12), and

two points together, the physics/psychology parallel implies that *drives* are no more to be taken as psychological atoms than the soul itself, and in principle every drive or affect is open to analysis that would reveal a complex internal structure composed of further drive- or affect-shaped substructures.[21]

Note the anti-reductionist consequence. The anti-atomist point—that the self is a complex multiplicity of psychological substructures—might have been thought (by a reductionist) to undermine the *reality* of a self independent from the drives, because such a self is just a collection, and collections are nothing over and above their members. But that simply cannot be *Nietzsche's* view, on pain of the same argument's likewise eliminating the reality of all drives whatsoever (*none*

Boscovich even compares them to Leibnizian monads (see Boscovich (1922 [1763]: 17, 35, 83, 113, and also Article 398). Moreover, he denies that there is a continuum of such real, material points (Boscovich 1922 [1763]: Articles 391, 393), and expressly countenances the hypothesis that there may even be *physically* indivisible *collections* of point-elements playing the role of extended atomic corpuscles (Boscovich 1922 [1763]: Articles 393, 398). (The idea is that the compound coheres due to attractive forces, and that the resulting cohesion is too strong to be broken by any physically possible repulsive force, because any repulsive force great enough would have to be located at such a distance from the collection that it would act on all its parts together, and so could not divide them.) Nietzsche apparently knew Boscovich in the original (he borrowed Boscovich's *Theoria philosophiæ naturalis* from the Basel library for four semesters running in 1873–5; see Crescenzi 1994), but the basic argument of *BGE* 12 shows that he fundamentally misunderstood these aspects of the theory. After all, without the continuum mechanical (mis)interpretation, Nietzsche's inferences in *BGE* 12 simply do not follow. Nietzsche clearly means his argument to deny that there is any simple, indivisible thing serving as the basic object of psychology. The reference to Boscovich was supposed to promote this conclusion by suggesting that such simple indivisibles have no credible explanatory role even in physical theory, which ought to be the best case for their use (see previous note, for the pattern of reasoning). Thus the analogy can go through only if Boscovich is (wrongly) taken to be offering a continuum theory of matter that rejects any 'particle-atom' (*BGE* 12) *in the specific sense* of simple, indivisible, and therefore indestructible physical particles. Probably Nietzsche misunderstood (or, in 1885–6, mis*remembered*) Boscovich because the composition of *BGE* 12 was guided by Lange's *Geschichte des Materialismus* (Lange 1902 [1873–5]). While Lange does not actually make the mistake Nietzsche does, he does encourage, or at least, suggest it by describing a quick and all-but-inevitable logical progression from Boscovich's denial that the atoms are *extended* to the fully dynamical, continuum mechanical view of force-centres, which he attributes to Faraday (see Lange 1902 [1873–5]: 192–3). Lange's complaints against materialists in this passage—(they unjustifiably hold onto the material atom just because it satisfies a 'need of the mind' for *sensible* objects, i.e. objects analogous to perceptible billiard balls and such, in physics)—clearly marks it as Nietzsche's proximal source (recall from *BGE* 12 Boscovich's 'triumph over the senses'!). I hypothesize that as he thought through the ideas of *BGE* 12, Nietzsche had Lange's account of anti-atomism and Boscovich in mind (or in front of him), and he simply did not bother to check whether Boscovich's actual theory was in fact analogous to his intended defense of anti-atomism in psychology. (Thanks to David Hills for extremely helpful discussion.)

[21] Richardson himself denies this consequence, and continues to treat (some) drives as atoms (see 1996: 44–5, *et passim*), but I do not see how the anti-atomist result can be avoided. After all, on the side of physics, anti-atomism is supposed to be a consequence of the basic conceptual structure of thinking in terms of forces rather than particles, and it is the identity of that conceptual structure across physical and psychological explanation that is supposed to underwrite the basic notion of will to power as an explanatory device proper to both domains. I defend a particular account of the kind of theoretical unification the will to power doctrine is supposed to provide in Anderson (1994). For a further, and somewhat independently motivated, defence of anti-atomism about the drives in the context of moral psychology specifically, see Anderson (2006).

of which are atoms). Ironically, the tendency to draw eliminativist or reductionist conclusions from the argument of *BGE* 12 turns out to be itself a symptom of the very 'atomistic need' Nietzsche criticizes, which appears here in the guise of a latent assumption that only the psychological *atoms* could be truly real!

A third observation about *BGE* 12 is that, even though the passage makes problems for transcendentalism, one might still have expected the naturalist interpreters of Nietzsche to have been more put off by its explicit treatment of naturalism itself. Not only does Nietzsche mention naturalists in a dismissive tone, but he also makes it rather clear that the position he would like to dismiss is precisely the kind of naturalist reading that concerns us—the view that there is nothing to the soul, or that 'the self' is in reality just some lower-level, more naturalistically respectable entity, like the material brain, or a bundle of impressions and ideas, or the strongest drive. To remind ourselves, while Nietzsche is keen to get rid of the soul *atomon* and the inference to immortality, 'Between ourselves, it is not at all necessary to get rid of "the soul" at the same time, and thus to renounce one of the most ancient and venerable hypotheses—as happens frequently to clumsy naturalists who can hardly touch on "the soul" without immediately losing it' (*BGE* 12). This third point thus reinforces the anti-eliminativist moral of the second. Nietzsche's agenda is to *change our conception* of the soul, not to get rid of it as an identifiable object of psychology over and above its subpersonal constituents.[22]

Fourthly, it is worth paying attention to the hypotheses about the soul that Nietzsche takes to be worth exploring. From the perspective of the atomism problem, the most important new 'soul-hypotheses' are the conceptions of the 'soul as subjective multiplicity' and of the 'soul as social structure of the drives and affects' (*BGE* 12). While emphasizing (against atomism) that the soul is something *complex*, both of these formulations tell against any strong eliminativism, or any reductionist position about the relation between the self and its constituents. After all, a social structure is something that goes beyond the individuals who participate in it—a more or less definite group reality that may or may not characterize those individuals and their relations.[23] Thus, the

[22] Of course, 'naturalism' is a term of remarkable plasticity, and what Nietzsche means to dismiss under the name 'naturalism' is probably different from, and possibly quite a bit cruder than, the naturalism advocated by his current-day interpreters. But as I argue in the text, the point crucial for our purposes is shared by both versions of naturalism. The 'clumsy naturalists' of *BGE* 12 presumably are—or at least *include*—popular mid-nineteenth-century German materialists who were determined to make the reductionist point that the soul can be nothing but an aggregation of matter (see Leiter 2002: 63–71). For the purposes of *atomism*, however, the key reductionist move is shared by a more current naturalist program (or interpretation) purporting to reduce the self to some aggregation of constituent attitudes (e.g. drives, affects), or a Humean 'bundle' of psychological states (be they impressions and ideas, or beliefs and desires). In both cases, the basic idea is to get rid of anything that deserves to be called a 'self', or 'soul', and Nietzsche's comment in *BGE* 12 clearly aims to *resist* that impulse. (Thanks to Elijah Millgram for discussion.)

[23] Or at least, so Nietzsche himself clearly believes. His commitment to the reality of social level phenomena is clearly on display, for example, in the *Genealogy*'s description of what was

social structure of drives and affects, though it admittedly incorporates the subpersonal attitudes and could not exist without them, is still presented as something more than just the drives and affects themselves. Likewise in the first formulation, Nietzsche presents the self not merely as a multiplicity of attitudes, but as a *subjective* multiplicity—that is, I take it, as a structure with the subjective capacity to inhabit attitudes of its own, including, potentially, attitudes towards its constituent drives and affects. Thus the specific hypotheses Nietzsche proposes about the soul tend to support the thought that the self has some emergent reality over and above its constituent drives and affects, and thereby to cut against eliminativist or reductionist naturalisms, just as *BGE* 12's cutting dismissal of naturalism would suggest.

Finally, I note in closing that Nietzsche presents the self not as identical to the strongest drive, nor as a bundle of drives, but as an ordered structure of drives *and affects*. It has been tempting for readers to take Nietzsche's frequent talk of 'drives and affects' together as pleonastic, such that 'affect' does not add anything to talk of 'drives'. But as our quick survey of the complexities of Nietzsche's moral psychological apparatus suggested, drives are not affects, and this assimilation is likely to be too quick. I will argue below that some real illumination can come from careful attention to their differences and interactions.

I concede that this reading of *BGE* 12 offers only a set of textual indications that Nietzsche (even in his anti-transcendentalist moments) accepted some notion of a self existing over and above its constituent attitudes. There is not yet any real *argument* showing how Nietzsche justified that commitment, why he needed it, or what philosophical work it does for him. In the next section, I will offer the beginnings of such an argument, based on a bit of (more-or-less) first-philosophizing about Nietzschean drives and affects, and their place within his larger moral psychology.

5. DRIVES *AND* AFFECTS: AN INITIAL FORAY INTO NIETZSCHEAN PSYCHOLOGY

The complexity of Nietzsche's moral psychology noted in Section 3 puts a quick end to any hope for a comprehensive treatment here. As an initial stab, I propose

accomplished by those who formed the first states: 'Their work is an instinctive creating of forms, impressing of forms; they are the most involuntary, unconscious artists there are:—where they appear, in a short time *something new appears* there, a ruling structure that *lives*, in which parts and functions are delimited and related to one another, in which nothing at all finds a place that has not first had placed into it a "meaning" with respect to the whole' (*GM* II, 17, first italics mine). Here, obviously, social organization has its own reality, separate from the individuals it organizes and depends on. Otherwise, there would be nothing 'new' to appear, with its own 'life', and the artists of state formation would not have introduced something new into the world.

to follow the suggestion of *BGE* 12 and focus on two of the most central attitudes: drives and affects.

Drives, affects, and their complements

I begin from an important result due to Paul Katsafanas (2008, ch. 4; and forthcoming). Based on a far-ranging and penetrating analysis of Nietzsche's (and wider nineteenth-century) uses of the closely related terms 'drive' and 'instinct', Katsafanas shows that Nietzschean drives are importantly different from desires. The crucial point concerns relative *complexity* with respect to the complements of the different attitude types. On one commonly assumed conception, desire takes a one-place complement: for example, I desire an object (that Burdick's chocolate truffle, say), or I desire some (propositionally structured) state of affairs (e.g. that I arrive home safely from a trip). Indeed, this simple, one-place structure contributes to the plausibility of counting desire as the fundamental conative attitude, out of which further attitudes with world-to-mind direction of fit should be constructed. But, so Katsafanas shows, drives take a *two-place* complement. A drive not only has a particular (propositional or individual) *object* that it tracks, but it also, and separately, pursues a more abstract *aim*—a characteristic pattern of activity of which the pursuit of this or that object of current attention is merely an instance.[24] For example, my drive for food can take any number of particular objects (e.g. the Burdick's chocolate truffle, the five-course meal I am in the midst of preparing, or simply that I am no longer hungry), but all these are merely particular occasions, suitably shaped for the object position, for the expression of the drive's broader aim, namely, the pattern of activity towards which it teleologically tunes my behaviour (in this case, eating).

To see the importance of this aim/object distinction, just reflect on the case where I am a compulsive eater: in such circumstances I cannot of course do without appropriate *objects* for my drive—indeed, seeking them is the main focus of my compulsive attention—but at the same time, no such objects actually satisfy me; as soon as I have eaten them, the drive reasserts itself (i.e. its pursuit of its *aim*), and I am off in search of a new object.[25] As Katsafanas nicely puts the point:

[24] Of course, this sort of aim/object distinction for drives gets substantial development in Freud's theory of drives, but Katsafanas shows that the same distinction is present and important throughout the tradition, going all the way back to the key early philosophical deployments of the notion of drives around the turn of the nineteenth century, e.g. in Fichte, Schiller, and Schopenhauer.

[25] From this point of view, it should be immediately apparent why drive psychology was so appropriate for *Schopenhauer's* purposes. It nicely generates a 'how-possible' explanation for the sort of futility of conation that is at the heart of his pessimism. According to that explanation, we are never satisfied, because what a drive *really* seeks is its aim, but all it can ever get is an object. Thus desiring reasserts itself (with its attendant suffering) almost as soon as it is satisfied. (Needless to say, not every form of drive psychology need be committed to this sort of pessimism about conation; the point is just that drive psychology crisply explains how the pessimistic theory is *possible.*)

> Drives are constant motivational forces that incline one to engage in certain activities or processes. Drives are not satisfied by the attainment of their objects, since their objects are just chance occasions for expression. In other words, the object serves as nothing more than an opportunity for the drive to express itself, by inclining the agent to engage in some activity or other. What the drive seeks is just this expression; the drive is satisfied[26] only when being expressed, when the process that it motivates is in progress. Accordingly, an activity that is motivated by a drive aims at the performance of the activity itself (Katsafanas 2008: 150).

Katsafanas goes on to outline a number of other important features of drives (they include or induce evaluative outlooks; they affect our perception of reasons (see Katsafanas, forthcoming: 31–2); they are continuous, diachronically persistent forces in a moral psychology; etc.). While I will make a bit of weather out of these characteristics, for the most part it will be enough for my purposes if we bear in mind the added complexity introduced by the observation that drives admit of the aim/object distinction.

What about affects? I think there a parallel point to be made for that other core Nietzschean attitude. 'Affekt' is a fairly common technical term in moral psychology. It refers to a class of attitudes that combine a passive, receptive responsiveness to the world with a reactive motivational output; these are states—standardly with a prominent *feeling* component—through which we detect the saliences of things and find ourselves motivated to respond. But even though it is a technical category, 'Affekt' tends to get ostensive rather than stipulative definition. As Janaway (2009: 52) observes, the affects Nietzsche talks about are very often inclinations or aversions, and at least the core paradigm affects are attitudes we nowadays think of as *emotions*: love, hate, anger, fear, resentment, joy, contempt, glorying, etc.

Like drives *sensu* Katsafanas, I submit, affect qua attitude takes (at least) a two-place complement.[27] In place of the aim/object structure characteristic of drives, affects are completed by a *stimulus object* and something like a *default* behavioural *response*. The attitude itself colours the salience and evaluation of the stimulus object and it governs both the pattern and the manner of the agent's default response.

[26] Note that for drives, at least for Nietzsche, being 'satisfied' and being 'activated' are not really distinct. This marks another fundamental difference between drives and desires, since drive *satisfaction* is something fundamentally different from desire satisfaction. (In normal cases, of course, when a desire is satisfied, it is extinguished.) Nietzsche makes use of this feature of drives to avoid the pessimistic inferences Schopenhauer derives from the moral psychology of satisfaction, discussed in the previous note.

[27] I introduce the qualification 'at least' because, in fact, recent discussions of the emotions suggest that matters are likely to be substantially more complicated. Emotions (at least often) have much more complex and ramified complements (see following note). But the two-factor complement structure I go on to identify in the text captures at least an important part of the story. It identifies a basic organizing structure exhibited by the complements of paradigmatic affects/emotions; still more complex complements can then be fitted into and/or around the two aspects I emphasize. Thus, the story below can serve as an adequate, albeit highly simplified, *idealization* for present purposes. (Thanks to Elijah Millgram for discussion.)

These three elements—stimulus object, default response, and the emotional 'colouring' of each—emerge clearly in paradigm cases of affect. For example, the affect of *ressentiment* is standardly activated by an appropriate stimulus object (another agent, or agent-like object, who does injury or stands athwart the agent's will), and it issues in a default tendency to respond by seeking revenge. The distinctive affective/emotional character of the attitude emerges both in the way it colours our perception of and attention to the stimulus object (recall here *ressentiment*'s perception of the noble man, 'but dyed in another colour, interpreted in another fashion, seen in another way by the venomous eye of *ressentiment*' (*GM* I, 11)), and also through the *manner* in which its evaluative framework shapes the pursuit of revenge (e.g. with 'hatred' of 'monstrous and uncanny proportions', 'the most spiritual and poisonous kind of hatred' (*GM* I, 7)—as opposed to vengefulness that 'consummates and exhausts itself in an immediate reaction, and therefore does not poison' (*GM* I, 10)). To take another example, the affect of joy will arise in response to some stimulus object (e.g. the long desired friend finally arrived, one's state of well-being, the fact that an enterprise has turned out successfully), and prompts a default expressive (re)action (an embrace, exaltation, celebration), where both the perception of the object and the manner of the reaction are governed by the distinctive emotional colouring of the affect. Or consider Nietzsche's frequent exploration of the affect of disgust, which for him so often takes the 'last man' as its stimulus object and proposes some cleansing or purifying reaction (recall the 'export' proposal of *GM* III, 26, among many other examples), all the while creating an evaluative perspective that governs both the perception of the stimulus (the 'hopelessly mediocre and insipid' 'maggot "man"' (*GM* I, 11, etc.) and the manner of the desired response (e.g. the spirit of Nietzsche's fantasized Anacreontic chair-kicking in *GM* III, 26).

I hope these few examples are sufficient to motivate the plausibility of a broadly two-factor account of affect complements. To sum up the point at the abstract, structural level, instead of taking a one-place complement such as perception (of an object) or desire (for an object), an affect is completed by both (a) some stimulus object that activates the affect, and (b) a default response upon which the affect primes us to act.[28] Finally, affects are like drives in that they come already 'evaluatively pre-loaded'. The feeling component of affect carries evaluative baggage

[28] I adopted this talk of emotions' 'priming us to act' from suggestions (in conversation) of Tim Bloser and David Hills. Again, the two-factor analysis I propose here is self-consciously offered as a simplifying idealization (see previous note), and I do not mean to deny that there may be further distinctions to be drawn as part of a fuller account of the structure of affect complements. For example, it has been proposed that we should distinguish the target object of an emotion (that towards which the emotion directs my thoughts and feelings) from its formal object (a relevant property ascribed to the target) and from the focus of the emotion. Many further distinctions have also been proposed in the literature. Here, I purport only to draw one fairly coarse-grained and general distinction meant to explain the peculiar combination of passive and active elements exhibited by affects/emotions.

that shapes and colours our perception of the stimulus and governs the manner characteristic of the default action path it suggests to us.[29]

The presence of two different types of complement helps to explain the curious combination of passive and active elements characteristic of affects. Affects seem to be essentially passive attitudes through which we are responsive to evaluatively salient features of the world, but at the same time fundamentally active *motivational* attitudes; as Janaway (2009: 52) points out, they are largely *inclinations and aversions*. Affects can play both roles because of their different complements: they show up as attitudes of passive sensitivity when we are focusing on their stimulus objects, but as motivational when we are focusing on the default action for which they prime us.

It is worth noting one final contrast to drives. The main object place of an affect is filled by its *stimulus* or cause, and is not necessarily the focus of the emotion or the target of the behaviour for which the affect primes us. Consider, for example, my fear of some danger threatening my loved one. Here the stimulus object (the danger) is distinct from the focus that orients the emotion (my loved one) and, in addition, it is not at all some target that I 'go for' in my fear-induced behaviour; on the contrary, I am trying to *flee* it or *block* it, and thereby teleologically pursuing some other goal like 'safety'. (Admittedly, we do attend to a feared stimulus object if it is specific enough, but precisely in order to get away from it. Perhaps the goal of our fearful behaviour is defined in terms of the stimulus, but *negatively*; we reach our goal when the object is gone.) For these reasons, affectively motivated action often seems relatively unfocused, or not under tight teleological control: my disgust at some spoiled food primes me to shove it away, but the impulse to fling it away or simply to close the door of the fridge as quickly as possible may well be a much *less* effective plan for removing it from the range of sensation than a behaviour that (temporarily) moves me *towards* it (e.g. opening the container and getting it all down the disposal).

Drives and affects working together

These structural observations cast interesting light on the *relation* between drives and affects, which turns out to be crucial to the main questions of this paper. The key points I will be emphasizing follow from the morphological features of drives and affects just canvassed, together with Nietzsche's anti-atomism, the evaluatively loaded character of both attitude types, and what I just called the unfocused character of affect-driven action.

[29] The fact that the *manner* of the default response is so directly governed by the feeling component of the affect is itself a compelling reason to insist that the response pattern really is part and parcel of the affect as an attitude, and does not arise from elsewhere (for example, from the affect's having recruited a separate drive, with which it acts in concert). Thanks to Elijah Millgram for pressing me to think through the motivations for this aspect of my analysis.

Note, first, the extent to which the general structural features of drives and affects tailor them to work together. Perhaps the most obvious source of this 'niceness of fit' is the difference between the targets, or pursuit objects, taken by drives and the stimulus objects taken by affects. By associating with an affect, a drive acquires sensitivity to a stimulus and thereby 'knows' when to activate; conversely, an affect can give better shape to its pattern of behavioural response by taking up a pursuit object from a drive.[30]

We can go further, however. In general, a drive represents its object and pursues its aim under the influence of some broad evaluative perspective, but for most drives the 'built-in' evaluative perspective proper to the drive itself is not sufficiently nuanced to explain the *range* drives exhibit in adjusting their expression to variation in the evaluative circumstances. To return to the Richardson example, my drive for food may always represent eating as a good, but I can eat lustily and with relish, or curiously, or sensuously, or with finicky particularlity, or dutifully under a 'food as fuel' mentality. This 'adverbial' variation—or anyway a great deal of it—is explained by the drive's *recruiting* an affect to further specify its evaluative perspective. Since the affect will have a prominent feeling component, it will add nuance to both the manner of the drive's aim-expression and its value-laden perception of its object. So, for instance, my drive for food might recruit the affect of greed and express itself gluttonously, or it might get caught up in my affect of despair or of slight disgust and express itself through a correspondingly inflected version of dutiful eating. Even better, think here of the way the presentational strategies adopted by a great restaurant conspire to slow us down and thereby induce more attentive eating that encourages a special focus on subtleties of flavour; tellingly, we call this 'setting a mood'.

A parallel point can be made for affects. As I noted, the 'unfocused' character of affect-driven action creates a natural opening for the affect to *recruit* a relevant drive to lend focus and firmer telic shape to the action for which it primes us. To take the most prominent Nietzschean example, the affect of *ressentiment*, under the right conditions, recruits the drive for power to hammer its vague impulse to get back against the happy into the incredibly subtle, highly structured, long-term, plan-shaped program of activity Nietzsche describes as the global revaluation of the noble pattern of values, or for short, the 'slave revolt in morality' (*GM* I; *BGE* 260, 262; *et passim*).

[30] In these and similar ways, the cooperative partnering of drives and affects suit them for roles in the sort of rationalizing explanations of behaviour with which we are familiar from the belief/desire folk psychology. That said, drives and affects *do not always* work together in this way. For example, drives need not partner with an affect so as to activate in environmentally appropriate circumstances, and in fact they *often* activate 'on their own', when circumstances are not especially appropriate. Thus, a drive psychology is particularly well suited to offer *non*-rationalizing explanations of behaviour that is not very rational. This sort of explanatory pattern (and its advantages) are well explored in Katsafanas' work on drives. (Thanks to Paul Katsafanas for illuminating discussion of this point.)

As the last example indicates, this kind of close interaction of drives and affects, based on mutual recruitability, is a basic and incredibly widespread feature of Nietzsche's actual explanations in moral psychology. In the interest of space, I won't try to discuss many specific instances; instead I will just gesture at three broad *patterns* of explanation that will be familiar. Consider first the force of Nietzsche's frequent classification of drives into the 'aggressive, form-giving' drives and 'reactive' ones (*GM* II, 12; I, 10–11; *et passim*). What separates the two classes? Among the important factors, as it seems to me, must be counted the characteristic differences in the affects they recruit to inflect their expression. Drives in the first class typically recruit the affect of aggression (or one of its many constituent affects or more specific versions); drives in the second class inflect themselves with *ressentiment* and its relatives, or else with a more general affect of reactivity.[31]

Second a similar point can be made about the explanatory strategy Nietzsche sums up with the observation, 'Regarding all aesthetic values I now avail myself of this main distinction: I ask in every instance, "is it hunger or superabundance that has here become creative?"' (*GS* 370). This question allows Nietzsche to separate the aesthetic drives he takes to be fruitful from destructive ones by appeal to the affects they tend to recruit, and also vice versa, to distinguish positive/affirmative affects from negative ones in terms of the drives they recruit. Thus Nietzsche insists that the artistic drive to destroy can be good (if it preferentially recruits 'Dionysian' affects of overflowing joy, hence expressing 'superabundance') or bad (if it tends to recruit vengeful affects and expresses 'hunger'). Likewise, the artistic drive to immortalize can be of the (good) type that recruits affects of gratitude or love (superabundance) or of the (bad) type, recruiting those of self-torture, e.g. in the case of Schopenhauerian 'romantic pessimism' (hunger). Conversely, the affects of gratitude and love themselves count as self-affirmative largely because they tend to recruit outwardly oriented drives of superabundance, whose aims conduce to the strengthening and integration of the agent and expanding the sphere of her power.[32]

A third class of cases involves the invidious distinction between 'natural' and 'unnatural' instincts (*GM* II, 24; *TI* V, 4–5; *et passim*). Nietzsche's official

[31] A related point might be made about a similar distinction between classes of drives in *BGE* 201; there Nietzsche separates a dangerous, aggressive, high-spirited class, characterized by affects tied to elation, the feeling of elevation, etc., from a class of drives promoting quiet, pro-social behaviour, which recruit affects related to timidity.

[32] It might seem tempting to reject any such analysis of affects, and instead take the distinction between affirmative affects such as gratitude and negative ones such as self-hatred as basic and irreducibly intuitive. After all, the positive affects do not (as it were) negate or attack the self. But the insufficiency of that simple, intuitive thought emerges quickly. For it is just not true that, for Nietzsche, all negative self-directed attitudes count as self-destructive like the ones he is trying to identify and relegate to this second (self-denigrating, hungry) class. Recall, for example, the importance for him of the inwardly directed affects crucially involved in *self-discipline*; contrary to intuitive appearances, those must surely also belong on the self-affirming side of the ledger for Nietzsche's purposes, despite their critical or even self-punishing attitude towards the self as it is.

position has to be that even the 'unnatural instincts' (to embrace the Beyond, etc.) are still expressions of some interest of life (see *GM* III, 13). They can count as 'against life' or 'anti-nature,' therefore, only in light of the configuration of the drive/affect interaction: natural drives are those that recruit self-affirming affects and unnatural instincts recruit affects of self-denial—*mutatis mutandis* for the case of life-affirming and life-negating affects.

I conclude that the drives and affects form a cross-hatched, mutually supporting structure of attitudes, whose integration rests on the way they are structurally tailored to recruit one another—e.g. with drives supplying a target object for affect-motivated action and affects supplying activation cues and also value-laden, nuanced specification to a drive's object perception and manner of expression. What follows from this picture?

The emergence of the (minimal) self

As a first consequence, consider that such a cross-hatched structure must routinely generate *one–many relations* between drives and affects. The entire explanatory apparatus depends on the availability of *the same affect* to be recruited to inflect the expression of many different drives. Think of all the different drives that recruit *ressentiment* to determine the manner of their expression—and the same goes for timidity, or joy, or hatred, or the affect of command. The same affect of love may be mobilized to modify the deployment of the erotic drive in one context, the artistic drive to immortalize (*GS* 370) in another, and the sociality drive in yet a third. Likewise, the *same drive* will often be recruited by many different affects. To cite the most central case, the will to power can be recruited by any number of affects to guide the pattern of their default response actions. The same goes for more specific first-order drives, such as erotic drive, which can enter to specify the responses of any number of affects: love, jealousy, fear, hope, curiosity, exuberance, and so on. For almost indefinitely complex treatment of the possibilities, recall Proust!

Perhaps surprisingly, this point yields a fairly strong implication for the Nietzschean self. If many drives can share the same affect, and many affects the same drive, then the drives and affects cannot be completely 'loose', 'distinct existences' in the sense made famous by Hume's 'bundle theory' of the self. If different drives depend for their own completion on being able to recruit one and the same affect, then they must be non-accidentally, functionally bound to each other in the same self, where that affect is available to be recruited. Similarly, different affects are bound to the self by their reliance on the recruitability of the same drive. The Nietzschean self is therefore not merely a Humean 'bundle' of instrinsically unrelated 'distinct existences', nor even a mere 'stage' upon which they enter and exit for one-off causal interactions. Instead, Nietzsche's conception of the relations between drives and affects forces the posit of a thicker notion of the self, existing as a *repository* of recruitable drives or affects that are always available to complete any of its given active drives or affects, such that (for example) *the same* affect of joy is ready to be recruited by my knowledge drive today, and my competitive drive tomorrow.

What is this 'repository self', presupposed by the one–many interactions of drives and affects? We can start by making clear what it is not. First, as we have just seen, such a self is not a mere aggregate, or 'bundle', of subpersonal attitudes, impressions and ideas. I hasten to concede that this minimal Nietzschean self is, in an important sense, built out of the drives, affects, and other attitudes, and could not be what it is without them. But the drives and affects could not be what they are without the whole Nietzschean self either, in that, for example, the typical complements and contents, and hence the functional capacities, of a given attitude will depend on which *other* drives and affects are available for it to recruit. Since the dependence relations between the self and its attitudes are *mutual*, the minimal self retains a real form of independence from the drives and affects. Moreover, even though the particular drives and affects are themselves standing attitudes that persist, rather than fleeting, occurrent states *à la* Hume, the minimal self must have its own *separate*, diachronic identity, which persists across *changes* of drives and affects. After all, the use of training or other forms of self-management to remove some drive or affect from the domain of recruitable attitudes (and the persistence of the self through the change) is a ubiquitous Nietzschean theme. Thus instead of a mere 'bundle' of individually fleeting attitudes, the minimal self is a diachronic, structured whole within which enduring drives and affects stand in causal and functional relations with identifiable patterns.

Second, however, it is equally important to emphasize—against various forms of Kantian transcendentalism—that the self in question is really *quite minimal*. When drives and affects recruit one another, the resulting patterns of relations among them (both causal and content/complement-based relations) emerge from the interactions of the drives and affects themselves; they are not relations (like that among the terms in a judgement) that would have to be established by an explicit or implicit act of 'synthesis' on the part of some unified agency separate from the drives. Moreover, the boundaries of the minimal self, unlike those claimed for the transcendental ego, are not identical with those of consciousness. In fact, the boundary mismatch obtains in both directions: the minimal self encompasses drives and affects it is not aware of, and it may have apparent conscious awareness of powers (e.g. the will) that are illusory. Thus, there can be no a priori argument from the alleged unity of consciousness to a strong, transcendental unity proper to the minimal self. In fact, the 'boundaries' of the minimal self are porous in principle; there is nothing to prevent my forming and acquiring new drives and affects, nor driving some of the ones I have out of existence. Finally, the degree of unity possessed by the minimal self is limited, not only in that drives and affects may be unavailable to central consciousness and completely non-transparent to one another, but also in that different constituents of the self may stand in oppositional, even quite conflictual, relations, resulting in weakness of will, and the like. Thus, the Nietzschean self as a whole is something over and above the constituent drives and affects, but it is not a simple, essentially unified and conscious, transcendental ego, which is fundamentally different in kind from the attitudes that compose it.

Is there dry ground to support such an intermediate position, between a Humean bundle and a Kantian transcendental self?[33] Peter Railton offers one reason to think there had better be, in work that inspired my talk of drives' and affects' 'recruiting' one another.[34] In several recent projects (Railton 2004; and this volume Ch. 2), Railton has been concerned to describe and defend a certain 'automaticity' that is ineluctably proper to action. As he notes, an action as simple as walking down the hall to get a drink of water inevitably involves a vast array of (in principle) identifiable sub-actions, sub-goals, responses, and adjustments—all of them guided by the environmental circumstances (via perceptual and kinaesthetic awareness) and by the overall goal set by the desire to drink, and all of them carried out intentionally and skilfully (in the mode Dreyfus calls 'skilful coping'), but utterly without explicit deliberation or even the formation of separate intentions. The last point is crucial for Railton's purposes. If we *did* have to form separate explicit, or even *implicit*, intentions or judgements about what we have reason to do in order to carry out each of these myriad intentional activities, we would be caught in an indefinite regress and action would never happen (Railton 2004). After all, each of those judgements or intention-formations would *also* be itself an action, which would require a prior judgement in its turn. Thus, it cannot be the case that some bit of activity cannot be mine, or cannot count as an action, unless I (i.e. a self distinct from the subpersonal attitudes and processes involved in the activity) separately endorse it, or intend it, or judge it to be good. Just as the 'regress of rules' argument demonstrates that there cannot be a rule for rule-following and thereby entails a basic capacity to apply a rule,[35] so analogously in the context of action the threat of regress demands that we recognize a prior and basic capacity to be aptly responsive to the circumstances, and (again on pain of regress) this capacity had better be a feature of our interacting subpersonal

[33] My colleague Allen Wood quipped, as a way of summing up the project of this paper, that its search for a middle way on these issues was most like trying to find a position to defend on dry land in the English Channel. I have little doubt that he will remain dissatisfied with the solution on offer, and would meet any riposte about my walking safely down the rue on the Isle of Guernsey by insisting that I am really well off the cliffs of Dover, and had better be a good swimmer!

[34] In his 2000 Kant lectures at Stanford, Railton adopted similar talk to explain the relation between reason and inclination in Kant, and that is where I first became aware of it. (For example, in moral motivation, reason first represents the good and then recruits a motivation to pursue it, whereas in non-moral motivation it is inclination that recruits a bit of instrumental reasoning to facilitate its pursuit of its object.) But Railton deploys this talk more generally in moral psychological theorizing (see, e.g. Railton 2004: 198).

[35] In his foundational version of the 'regress of rules' argument, Kant identified this basic capacity to apply a rule as the power of *judgement*, see *Critique of Pure Reason*, A 132–4/B 171–4.

attitudes (beliefs and desires) themselves and not something exercised by a separate, central agency:

> Belief and desire can operate without regress to yield intention if intentions can form and operate 'automatically'... through a kind of self-organization around ideas. Just as molecules with a certain architecture and composition can *crystallize*... without needing any guiding hand, so beliefs and desires with the right architecture and composition can crystallize into action-guiding intentions by clustering around an idea... without any guiding hand. Indeed, any sort of a *guiding* hand shaping the process of intention formation would itself have to be an intentional process. Agency, then, also confronts a regress problem... It had better be possible for intentions to emerge without being intended, their formation guided directly by beliefs and desires themselves (Railton 2004: 198).

When our attitudes potentially 'crystallize' in this way, they come together in a self that forms 'a structured, functional whole', and not just a Humean bundle (Railton 2004: 200).

In the Nietzschean minimal self, drives and affects are self-organizing in very much this sense. This possibility should not be surprising. As Richardson already noted, it is a basic feature of Nietzschean drives that they can combine to form larger units, in the relations he calls drive 'mastery' and 'tyranny' (Richardson 1996: 32–5, *et passim*). What we are now in a position to see, however, is that such combinations are only the beginning of the story. Not only can drives combine to form more complex drives, and not only can our attitudes coalesce (or 'crystallize') into strictly looser structures around particular intentions and patterns of action along the lines sketched by Railton, but further, there is a still looser whole into which the standing drives and affects organize themselves for the purposes of recruiting one another to secure their contents and complements. This larger, looser structure is the minimal self, a functional grouping of drives and affects that permits such mutual recruitability.[36]

Given Nietzsche's general anti-atomism and his views about drive/affect interaction, it makes sense to treat each of the things contributing to the self—i.e. each drive, affect, higher-order attitude, etc., up to and including the self as a whole—as a psychological object in its own right, even though they all stand in relations of mutual dependence. The minimal self is but one psychological structure among the others. It acquires the right to the name 'self' simply in virtue of being the emergent structure that encompasses *all* of the substructures

[36] Coalescing around a particular action or intention is a 'strictly looser' (self-)organization than drive mastery, since it is an occasional and repeatable (if temporally extended) cooperation among drives and affects, which remain distinct standing attitudes with their own characteristics. The drives and affects involved would normally exist, complete with their own life and effects within the self, both before and after the 'crystalization' event(s). By contrast, in drive mastery, one drive subsumes another, which loses its separate identity and has its defining aim reshaped by the new whole. The minimal self is a still looser whole, in that its constituents are not interrelated by their having been (actually) recruited by one another, but by their mutual availability for (possible) recruitment.

available for recruitment by one another.[37] It does also thereby gain a distinctive relation to the constituent psychological structures, based on the very looseness of its internal organization. Unlike a mastering drive, or even a 'crystallized' complex of drives and affects, the self—qua the emergent structure encompassing *all* the co-recruitable attitudes—can suffer from a 'gap' between its own activity and that of some constituent(s). Just because an attitude is recruit*able*, it does not follow that it will successfully be recruited in the appropriate circumstances. But such a recalcitrant drive or affect remains part of the totality, since it can still activate itself on its own and recruit (or be recruited by) the self's other attitudes. In this sense, the minimal self can remain 'responsible' for a recalcitrant attitude as something that belongs to it—by contrast to a mastering drive or a 'crystallization', wherein any attitude that is not presently and actually functionally integrated is simply not a part of the emergent whole, but a separate factor.

Suppose, then, we have successfully identified an emergent, complex psychological object built out of the constituent attitudes. Still, does Nietzsche have any right to think of such an object as a *self*, as a '*subjective* multiplicity' (*BGE* 12)? Some significant evidence in Nietzsche's favour on this point emerges from consideration of overarching *moods*. More or less global moods such as *depression* or standing (as opposed to occurrent) *joy* are best understood, I submit, as higher-order affects. They involve standing dispositions for some first-order affect (or characteristic range of affects) to be activated, coupled with a systematic attention- and sensitivity-bias towards the stimulus objects appropriate to those affects. But moods are not *merely* dispositions of first-order affects to be activated. A mood is also itself a particular (higher-order) attitude, which represents the world *and* the other affects within the self as being a certain way. Even though my mood may not be a sharply defined self-conscious attitude expressly owned by a unified 'I'—after all, I can be strongly in the grip of a mood without even being consciously aware of it—still, the mood operates as a kind of collective condition within which my other attitudes have to operate and with which they have to contend—a kind of 'weather system' influencing my other attitudes. Because of its global character and its self-referential features as a higher-order attitude, a

[37] Elijah Millgram (personal correspondence) offers the intriguing objection that, on Nietzschean grounds, this move ought to be insufficient to delineate the self, since Nietzschean drives are always seeking mastery over one another without any discrimination between potential targets of appropriation 'inside' and those 'outside' the 'self'. To put it colourfully, for Nietzsche the drives are always trying to 'eat the world' and so there is no usable sense available of 'all' the drives and affects that make up myself. I think a view like mine should concede that the boundaries of the Nietzschean self are not fixed in advance in some permanent, principled way; indeed, this was one of the features we saw distinguishing it from Kantian, transcendentalist conceptions above. That said, at any given time, there will be a (more or less) clear answer to which elements belong within my self, based on which ones are in fact potentially available for easy recruitment. If some new drive or affect later becomes a recruitable participant in the self's activity, then the self has expanded to encompass a new element. (I believe that some fuzziness around the edges and ambiguity about borderline cases is tolerable, here, and indeed should count as a feature, not a bug, from the point of view of Nietzsche interpretation, but a fuller discussion will have to wait for another occasion.)

mood like depression or joy counts as an attitude inhabited by *the whole minimal self* and not just an outgrowth of some particular constituent drive or affect. For just that reason, Nietzsche places heavy emphasis on mood-like higher-order affective responses when, as in the thought experiment of eternal recurrence, he looks for indicators relevant to the evaluative assessment of the whole self, or individual life. But now, given such higher-order affects, we can say in a serious way that the Nietzschean minimal self is a genuinely '*subjective* multiplicity' (*BGE* 12, my ital.)—a self that inhabits attitudes of its own, including ones directed at itself.

We now have everything we need to provide a preliminary answer to our main question. Nietzsche's moral psychology provides materials for, and indeed forces him to postulate, a self that is something over and above its constituent drives and affects. Moreover, despite remaining fairly minimal, the self so understood does have the capacity to take up attitudes (including evaluative attitudes) towards the world and also towards itself and its drives and affects. These reflexive attitudes may include consciously reflective or even deliberative attitudes such as the control of affective interpretations involved in perspectivist objectivity or the more or less explicit attitudes of self-management involved in Nietzschean self-overcoming, self-mastery, and so on. But as we have just seen, they can also take the form of moods and comparable higher-order attitudes, which *lack* any such reflective, deliberative character. For just that reason, the postulation of the *minimal* self is warranted even for agents who lack the more deliberate or reflective reflexive attitudes (e.g. because they are catastrophically weak-willed, deeply divided against themselves, etc.). Not only slaves, Christians, and ascetics, but even those chaotic wantons 'who stand in dire need of being ascetics' (*TI* V, 2) still have a minimal self, separate from the drives and capable of expressing telling attitudes towards them, attitudes which Nietzsche takes to be symptomatic indicators of the value those selves manage to instantiate.

6. CONCLUSION: THE NORMATIVE CONCEPTION OF THE SELF, OR SELFHOOD AS A TASK

My aim was to work out some details of Nietzsche's moral psychology, and thereby to assess the prospects for a conception of selfhood that is genuinely Nietzschean, but also plausibly possessed of one distinctively Kantian faculty: the capacity to 'stand back' from one's own attitudes and assess them. This capacity was of particular interest because it makes possible autonomy, a value whose importance Nietzsche often seems to endorse right along with Kant. In conclusion, I should make at least a gesture in the direction of connecting our results about the self to the larger issues about autonomy.

I have argued that Nietzsche rightfully posits a minimal self possessing evaluative attitudes about its drives and affects, and perhaps even a self-conception. But while this might be a capacity *needed for* anything like autonomy, it certainly falls far short of *achieved* autonomy. As we just saw, even desperately weak-willed individuals who are wholly at the mercy of their drives—that is, people who are deeply unfree and certainly incapable of *ruling themselves* autonomously—have a minimal self with this capacity for reflexive self-assessment. In fact, if they lacked that capacity, we could not understand them as *weak-willed* at all; the drive that actually determined their behaviour would *ipso facto* count as the ruling drive (and therefore as their self in the only meaningful sense), and there would be no sense in which the agent/wanton was acting against her own values, will, or considered assessment.

(Let this count as one final broad-brush textual reason for rejecting extreme naturalism. Criticism of weakness of will and related forms of evaluative inconsistency are central to Nietzsche's core philosophical stances, including the key arguments of the critique of Christianity. Eliminative naturalism about the self lacks the resources to make sense of these complaints; hence the reading is not adequate to Nietzsche's purposes.)

But now, if the minimal self with its capacity to stand back from the drives is insufficient for autonomy, where does that leave Nietzsche's apparent valuation of autonomy? Is *that* notion, and/or whatever notion of selfhood is needed to underwrite it, still loaded with 'moral excess' and therefore in need of Williams-style purification? I think not, and we can see why by returning to the normative conception of selfhood as a task or achievement.

As I noted above, Nehamas (1985) and Schacht (1983), followed by several others more recently,[38] all observe that Nietzsche often deploys the concept of selfhood not to capture some descriptive structure or property of a person's moral psychology, but instead as a *norm*, thereby treating selfhood as a kind of *task* that is set for us, or an *achievement* made by some people but not others. For example, such a conception is needed to underwrite Nietzsche's ideal of self-creation, which gains typical expression in his famous praise of Goethe:

> 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]

The notion of self-creation deployed here is superficially paradoxical: the activity in question could not be *self*-creation unless one did it oneself, but that very self (namely, oneself) is the thing that is supposed to be created, and thus should first come into existence *only through* the process. Obviously, Nietzsche

[38] Notable treatments I am aware of include Gemes (2009) and Janaway (2009). I/we make similar suggestions in Anderson (2006), and earlier (in a version that closely follows Nehamas) in Anderson and Landy (2001).

assumes that Goethe was already some kind of self before he 'disciplined himself to wholeness'; indeed he was *him*self, in a sense sufficient for the self-disciplining activity to count as his own. But Goethe became *more truly* himself—he realized his selfhood in some stronger sense—by attaining the wholeness he sought, and it is this truer self that he '*created*'. The paradox is dissolved, therefore, by a distinction between two conceptions of selfhood: one descriptive conception that includes the moral psychological capacity of the person to frame and carry out the plan of self-creation (or any other plan), and a second, normative conception of the 'true self', which encapsulates the ideal being pursued.

The same normative sense of selfhood is also in play in Nietzsche's ubiquitous praise of genuine or 'strong' individuals, most famously in his encomium to the 'sovereign individual':

> If we place ourselves at the end of this tremendous process, where the tree at last brings forth fruit, where society and its morality of custom at last bring to light *to what* they have been only the means: then we will find as the ripest fruit on its tree the *sovereign individual*, like only to himself, liberated again from morality and custom, autonomous and supramoral (for 'autonomous' and 'moral' are mutually exclusive), in short, the human being with his own independent, long will, who is *permitted to promise*—and in him a proud consciousness, quivering in every muscle, of *what* has finally been achieved and made flesh in him, a real consciousness of power and freedom, a feeling of the completion of man in general. This emancipated one, who really *may* promise, this master of a free will, this sovereign—how should he not be aware of his superiority . . . ? (*GM* II, 2)

Here, clearly, individuality is not merely a thin, descriptive property possessed automatically by every single human being; on the contrary, it is a rare and high *achievement*, attained by a few especially great people at the cost of the sacrifice of untold ordinary mortals who are not even *aware* of the kind of greatness exemplified by those special individuals.

Tellingly, in both these cases Nietzsche tightly ties the normative conception of selfhood, or individuality, to the value of autonomy. Genuine selves realize *that* value: by creating himself, Goethe emerges from self-creation as 'a spirit who has *become free*' (*TI* IX, 49); the sovereign individual is 'autonomous' and 'liberated from custom'.

In my view, the connection Nietzsche wants to find between self-creation and autonomy, and indeed his conception of autonomy itself, finds a natural moral psychological basis in the distinction between this normative conception of selfhood and the minimal self. The minimal self *is just* a certain moral psychological structure among the drives and affects, no matter how conflictual and disunified they may be. One must attain something further to become a self in the stronger, normative sense. I have argued elsewhere (Anderson 2006) that Nietzsche operationalizes the relevant norms largely via appeals to *strength*, where strength is understood, in turn, as strength of will (as opposed to weakness of will) and thus in terms of a certain kind of *unity*, or greater integration, among the drives and affects. So much is clearly envisioned in the description of Goethe's achievement, for example. But what makes such unity count as *one's own* is precisely its having been *self-generated*—that is, the unity among my drives and affects arises from regulating control over them that is exercised by and through the attitudes proper to the emerging self: to be noble is 'to have and not have one's affects . . . at will; to condescend to them . . . to make use of [them] . . . ' (*BGE* 284). In just such circumstances, attaining the normative self counts as self-creation, and it also realizes a recognizable form of autonomy. The self here follows values and laws it gives to itself.

But this stronger conception of autonomous selfhood, no matter how normative it is and however far it outstrips minimal selfhood, is no more plausible a target for a Williams-style critique of 'moral excess' than is Nietzsche's complex moral psychological apparatus. For even when achieved, autonomous selfhood is not anything fundamentally different in (psychological) kind from the minimal self: the normatively ideal self is still a structure of drives and affects; it is just a more unified, more harmoniously ordered, more internally disciplined and effective 'social structure' or 'subjective multiplicity'—one last time, it was '*totality*' that Goethe wanted; 'he disciplined himself to wholeness' (*TI* IX, 49). As far as I can tell, Nietzsche adopts an 'empiricist' attitude towards normative selfhood, in the following sense. He is *not* claiming that there must be some special, morally relevant psychological faculty in all persons which automatically suits them a priori to be targets of his evaluative judgement about whether they are autonomous. On the contrary, he merely articulates an ideal for the relation that ought to obtain among whatever drives and affects we happen to have. Whether any individual person attains that ideal or not is an empirical question, to be settled by the best interpretation of the person's life. We may dispute with Nietzsche about the suitability of his ideal, fair enough. But the psychology it relies upon remains innocent of any suspicion of 'excess moral content', precisely because the relevant notion of selfhood is not a fact but a norm—either someone exemplifies it (in which case its reality is conceded) or not (in which case Nietzsche's theory never claims that autonomy, or indeed any self in the normative sense, was present in the first place).

To conclude, neither transcendentalist nor naturalist readings can satisfactorily account for Nietzsche's conception of the self. Nietzsche need not endorse a transcendental role for the unified consciousness, for his moral psychology affords materials sufficient to explain how a self over and above the various drives and affects can emerge from the interactions of the drives and affects themselves. Such a self is essentially complex and not co-extensive with consciousness, so it does not carry the strong properties of the transcendental ego with which readers like Gardner would saddle the Nietzschean self. At the same time, even this internally complex, minimal self is something over and above its constituent attitudes, so naturalistic reduction or eliminativism about the self is equally inadequate. The insufficiency of such readings for Nietzsche's purposes is especially glaring when we turn to the autonomous self he idealizes, which exhibits a stronger, self-generated form of unity that far outstrips a mere 'bundle' of drives and affects.

Of course, a determined naturalist might simply try to deny that Nietzsche intends to repose any such value in the self, but such a position increasingly strikes me as incredible. The self's relation to itself and its attitudes towards itself ground the central normative judgements of Nietzsche's philosophy, a fact underlined by the powerful recent strand of readings advocating essentially 'practical' interpretations of the eternal recurrence doctrine,[39] as well as by attention to core Nietzschean concerns such as the creation of values, self-overcoming, and self-mastery. Even the urgency of Nietzsche's hope for 'new philosophers' rests on the same valuation of reflective self-control; they are important precisely because they will 'teach man the future of man as his *will*, as dependent on a human will' (*BGE* 203; emphasis in original). While such self-control *can* threaten to turn ascetic if overdeveloped (see *GS* 305), it nevertheless remains, when suitably deployed to promote the self's autonomy, absolutely central to Nietzsche's conception of the good life:

> A free human being can be good as well as evil, but . . . the unfree human being is a blemish upon nature and has no share in any heavenly or earthly comfort . . . [and] everyone who wishes to become free must become free through his own endeavor . . . [for] freedom does not fall into any man's lap as a miraculous gift (*UM* IV, 11; quoted in *GS* 99).

BIBLIOGRAPHY

In Nietzsche

For Nietzsche's German, I used *KSA*. I also made use of the following translations, cited by abbreviations. Date of first publication appears at the end of each reference. Parenthetical citations in the text refer to Nietzsche's section numbers, which are the same in all editions.

Beyond Good and Evil, trans. Walter Kaufmann. New York: Vintage, 1966.
Daybreak, trans. R. J. Hollingdale. Cambridge: Cambridge University Press, 1982.
On the Genealogy of Morality, trans. M. Clark and A. Swensen. Indianapolis: Hackett, 1998.
The Antichrist, trans. Walter Kaufmann. New York: Viking, 1954.
The Gay Science, trans. Walter Kaufmann. New York: Vintage, 1974.
Twilight of the Idols, trans. Walter Kaufmann. New York: Viking, 1954.

[39] Here, see especially Clark (1990: 245–86) and Reginster (2006: 201–27, *et passim*), but there are many other important contributors, including Löwith (1997 [1935]), Soll (1973), and Nehamas (1985). I advocate a minimal version of this broadly practical interpretation in Anderson (2005a: 196–203), and Anderson (2009).

Untimely Meditations, trans. R. J. Hollingdale. Cambridge: Cambridge University Press, 1983.

Other sources

Anderson, R. Lanier (1994). 'Nietzsche's will to power as a doctrine of the unity of science', *Studies in History and Philosophy of Science* 25: 729–50. Reprinted in *Angelaki* (Special Issue: Continental Philosophy and the Sciences: the German Tradition, ed. Damian Veal) 10 (2005): 77–93.

——(2005a). 'Nietzsche on truth, illusion, and redemption', *The European Journal of Philosophy* 13: 185–225.

——(2005b). 'Neo-Kantianism and the roots of anti-psychologism', *The British Journal for the History of Philosophy* 13: 287–323.

——(2006). 'Nietzsche on strength, self-knowledge, and achieving individuality', *International Studies in Philosophy* 38: 89–115.

——(2009). 'Nietzsche on redemption and transfiguration', in Landy and Saler (eds), (2009: 225–58).

——and Landy, Joshua (2001). 'Philosophy as self-fashioning: Alexander Nehamas's Art of Living', *Diacritics* 31: 25–54.

Avenarius, Richard (1888). *Kritik der reinen Erfahrung*. Leipzig: Fues.

Boscovich, Ruggero Giuseppe (1922). *A Theory of Natural Philosophy, Put Forward and Explained by Roger Joseph Boscovich, S.J. Latin–English Edition from the Text of the First Venetian Edition Published under the Personal Superintendence of the Author in 1763, with a Short Life of Boscovich*. Chicago: Open Court.

Clark, Maudemarie (1990). *Nietzsche on Truth and Philosophy*. Cambridge: Cambridge University Press.

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

Crescenzi, Luca (1994). 'Verzeichnis der von Nietzsche aus der Universitätsbibliothek in Basel entliehenen Bücher (1869–1879)', *Nietzsche-Studien* 23: 388–442.

Gardner, Sebastian (2009). 'Nietzsche, the self, and the disunity of philosophical reason', in Gemes and May (2009).

Gemes, Ken (2006). '"We are of necessity strangers to ourselves": the key message of Nietzsche's *Genealogy*', in C. Acampora (ed.), *Nietzsche's Genealogy of Morals: Critical Essays*. Lanham, MD: Rowman & Littlefield.

——(2009). 'Nietzsche on Free Will, Autonomy, and the Sovereign Individual', in Gemes and May (2009).

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

Hatfield, Gary (1991). *The Natural and the Normative*. Cambridge, MA: MIT Press.

Hill, Kevin (2003). *Nietzshce's Critiques: the Kantian Foundations of his Thought*. Oxford: Oxford University Press.

Janaway, Christopher (2009). 'Autonomy, affect, and the self in Nietzsche's project of Genealogy', in Gemes and May (2009).

Kant, Immanuel (1997 [1781/1787 = A/B]). *Critique of Pure Reason*, trans. P. Guyer and A. Wood. Cambridge: Cambridge University Press. Citations are to the pagination of the first (A=1781) and second (B=1787) editions.

Katsafanas, Paul (2008). *Practical Reason and the Structure of Reflective Agency*. Ph.D. Diss., Harvard University. Cambridge, MA. Available online at www.unm.edu/~katsafan.

——(forthcoming). 'Nietzsche's philosophical psychology', in Richardson and Gemes (forthcoming).

Korsgaard, Christine (1996). 'Personal identity and the unity of agency: A Kantian response to Parfit', in *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.

Landy, Joshua, and Saler, Michael (eds). (2009). *The Re-Enchantment of the World: Secular Magic in a Rational Age*. Stanford, CA: Stanford University Press.

Lange, F. A. (1902 [1873–5]). *Geschichte des Materialismus und Kritik seiner Bedeutung in der Gegenwart*, 7th edn., following the 2nd, rev. edn. Ed. and Intro., H. Cohen. Leipzig: J. Baedeker.

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

——(2009). 'Nietzsche's theory of the will', in Gemes and May (2009).

——and Knobe, Joshua (2007). 'The case for Nietzschean moral psychology', in Leiter and Sinhababu (2007: 83–109).

——and Sinhababu, Neil (eds) (2007). *Nietzsche and Morality*. Oxford: Oxford University Press.

Löwith, Karl (1997). *Nietzsche's Philosophy of the Eternal Recurrence*, trans. J. H. Lomax. Berkeley: University of California Press.

Mach, Ernst (1910). *Contributions to the Analysis of the Sensations*, trans. C. M. Williams. Chicago, IL: Open Court.

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

Railton, Peter (2004). 'How to engage reason: the problem of regress', in Wallace *et al.* (2004: 176–201).

Reginster, Bernard (2003). 'What is a free spirit? Nietzsche on fanaticism', *Archiv für Geschichte der Philosophie* 85: 51–85.

——(2006). *The Affirmation of Life: Nietzsche on Overcoming Nihilism*. Cambridge, MA: Harvard University Press.

——(2012). 'Autonomy and the self as the basis of morality', in Allen Wood (ed.), *Cambridge History of Philosophy in the Nineteenth Century*. Cambridge University Press.

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

——and Gemes, Ken (eds) (forthcoming). *The Oxford Handbook of Nietzsche*. Oxford: Oxford University Press.

Risse, Matthias (2007). 'Nietzschean "animal psychology" versus Kantian ethics', in Leiter and Sinhababu (2007: 53–82).

Smith, Michael (1994). *The Moral Problem*. Oxford: Blackwell Press.

Soll, Ivan (1973). 'Reflections on recurrence', in R. Solomon (ed.), *Nietzsche: a Collection of Critical Essays*. Garden City, NY: Anchor Doubleday, 322–42.

Wallace, R. Jay, Pettit, Philip, Scheffler, Samuel, and Smith, Michael (eds) (2004). *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*. Oxford: Oxford University Press.

Williams, Bernard (2006 [1993]). 'Nietzsche's minimalist moral psychology', in Williams, *The Sense of the Past: Essays in the History of Philosophy*, ed. M. Burnyeat. Princeton, NJ: Princeton University Press, 299–310. First published in *European Journal of Philosophy* 1: 4–14.