Agency and the Foundations of Ethics: Nietzschean Constitutivism

Published: October 16, 2013

Paul Katsafanas

Paul Katsafanas, *Agency and the Foundations of Ethics: Nietzschean Constitutivism*, Oxford University Press, 2013, 267pp., $75.00 (hbk), ISBN 9780199645077.

Reviewed by Alex Silk, University of Birmingham

If your prior credence that a book subtitled *Nietzschean Constitutivism* would be intelligible isn't high, you're probably not alone. This makes Paul Katsafanas's book all the more impressive and important. Katsafanas offers a comprehensive examination of constitutivism in ethics, including a lucid exposition and critical discussion of previous constitutivist theories, as well as a novel version of constitutivism that draws on developments in previously untapped areas -- most notably, Nietzsche's ethics, metaethics, and philosophical psychology. Chapters 1 and 2 examine constitutivism in general, motivating the view, explaining what is essential to it, and defending it against some of the most pressing objections; Chapters 3 and 4 examine two of the most prominent existing constitutivist accounts, those of David Velleman and Christine Korsgaard; and Chapters 5 through 9 develop Katsafanas's positive version of constitutivism: Nietzschean Constitutivism. The writing is consistently clear, the argumentation reliably rigorous: sections are bookended by a roadmap and recap, arguments are expressed in (something close to) premise-conclusion form, and possible responses to objections are made explicit up front; one rarely finds oneself lost in the dialectic. The book is a valuable read not only for

those interested in constitutivism, but also for anyone with a serious interest in ethical theory, philosophy of action, moral psychology, and Nietzsche studies more broadly.

The book is not without its flaws -- though, to Katsafanas's credit, it is often the characteristic presence of the above virtues that makes salient their absence in discussions of certain critical issues. Among the book's most important contributions is that it highlights a putative pervasive problem facing existing versions of constitutivism. It is this that elucidates the structure of a successful constitutivist theory -- or at least a constitutivist theory that stands the best chance of succeeding -- and thus paves the way for the book's primary positive project of developing such a theory in detail. However, a cost of the impressive range of Katsafanas's discussion is that many important pieces of his positive view are only sketched. Objections to alternative accounts are sometimes accepted prematurely, and analogous worries for his theory are sometimes hastily dismissed. In this review I will focus on examining, first, whether previous constitutivist theories are in fact vulnerable to Katsafanas's charge, and, second, whether Katsafanas's Nietzschean Constitutivism fares better. I will close by briefly considering two issues -- one philosophical, one interpretive -- concerning what is Nietzschean in Nietzschean Constitutivism: the thesis that power is a constitutive aim of action.

Constitutivism attempts to ground normative facts -- facts about what is valuable, what there is reason to do, etc. -- in facts about what is constitutive of agency or action. Constitutivists agree that aims generate normative reasons, reasons to satisfy those aims. And they agree that action itself has a constitutive aim, i.e., that every token action A aims at some goal, where having this aim is part of what constitutes A's being an action (39). (Some constitutivists, like Korsgaard, couch their theories in terms of constitutive principles, but I will bracket this complication here.) Where constitutivists differ is on what this constitutive aim is, and thus on what (universal) reasons are generated by this aim. Katsafanas's strategy is

to grant Velleman and Korsgaard their conceptions of action but deny that their theories of practical reason -- their theories of what normative reasons are generated by the constitutive features of action -- follow.

Katsafanas's summary of his objection is worth quoting in full:

The constitutivist seeks to show that actions have some feature, F, which both *makes them actions* and *makes them good actions*. In order to account for the possibility of bad action [an action which there is less reason to perform than another available action], the constitutivist needs to open a gap between actions *without* feature F (the good-making feature) and actions *with* feature F. The only way to do this . . . is to claim that F comes in degrees . . . The constitutivist can then claim that good actions manifest a high degree of F, whereas bad actions manifest lesser degrees of F. However, the constitutivist then faces the question of why we should aim at manifesting actions with the highest degree of feature F, given that we could perform actions with low degrees of F without thereby ceasing to act. . . . Rather than establishing merely that we aim at manifesting some degree of F, the constitutivist needs to establish that, in every action, we aim at manifesting the *highest degree* of F. Neither Korsgaard nor Velleman succeeds in showing this. (107-108; cf. 65, 109)

Consider Velleman's theory, on which action's constitutive aim is self-knowledge, or knowing what you are doing. If this is right, then we have reason to perform actions that generate some degree of self-knowledge. The worry is that nothing in the theory captures why we have more reason to perform actions that more fully satisfy action's constitutive aim and generate more self-knowledge. This problem generalizes to all constitutivist theories (78n.13, 107-108).

The solution, Katsafanas claims, is to employ a "differentially realizable" constitutive aim -- an aim that can be fulfilled to different degrees (e.g., an aim of earning money (75)). Katsafanas sometimes objects that Velleman and Korsgaard do not employ a differentially realizable aim and fail to

distinguish differentially realizable and non-differentially realizable ("simple") aims (see esp. 101, 109). But these claims are overstated. Katsafanas's real objection is that the arguments Velleman and Korsgaard give for their theories do not establish that the constitutive aim of action is differentially realizable. The discussion of these issues is rich. For reasons of space I will consider just one of Katsafanas's objections: his objection to Velleman's appeal to improvisational actors.

Katsafanas considers an improvisational actor impersonating a mobster torn between a desire to avenge his lover and a desire to avoid a conflict that could undermine his business. He claims that it would be "implausible . . . that the improvisational actor is somehow a failure for choosing [to forgo revenge]" (82), even if seeking revenge would make most sense for the character. After all, the actor may have other goals, for example, surprising the audience. But "if this is true of the improvisational actor," Katsafanas objects, "it should also be true of the ordinary agent" (82). However, this is where the analogy seems to break down. If the real-life mobster forgoes revenge even though he knows that this makes less sense, given his tempestuous nature, then there *is* a sense in which his action is deficient. The action seems deficient in precisely the same way that Katsafanas grants that a behavior that isn't intelligible at all is deficient (73-74). Since Katsafanas grants Velleman's theory of action for the sake of argument, he ought to grant that the mobster's action is deficient *qua action* (even if there is another respect in which the action is not deficient that renders it what he ought to do).

There is also a more general worry with Katsafanas's critiques of previous constitutivist theories. He insists (e.g., 107) that if the constitutivist treats the relevant constitutive aim G as differentially realizable, she must establish that action constitutively aims at the maximal degree of G; for if she only establishes that action constitutively aims at some degree of G, she won't capture conclusions about weights of reasons. But this fails to appreciate what is involved in treating an aim as gradable. It forces the

constitutivist's theory back into the model of simple aims: the aim of attaining a minimal/maximal degree of G is itself a simple aim; one either satisfies it or not. The aim of a horror film needn't be to be maximally scary (whatever that would mean). The aim is simply *scariness*, and the more scariness the better. Aiming at a gradable property G is not equivalent to aiming at the maximal (/minimal) degree of G. Katsafanas elucidates how Velleman and Korsgaard ought to explicate their theories, though they may have more resources to do so than he assumes.

Given the extent to which Katsafanas presses his objections to Velleman's and Korsgaard's theories, it is surprising how cursory his remarks are concerning how these objections might carry over to his own account. According to Katsafanas's Nietzschean version of constitutivism, action has two constitutive aims: (i) *agential activity*, understood as approval of one's action given complete information about its etiology, and (ii) *power*, understood not as brute domination but as "seeking and overcoming resistance to one's ends" (159). He appeals to both of these constitutive aims in his primary argument that agents are committed to aiming at maximal power (207-209). The argument appears to be something like this:

1. In performing any action A, one aims at (commits to) endorsing A given full information about A's etiology.
2. The etiology of every action includes will to power.
3. So, "Merely in virtue of acting, we become committed to approving of will to power. . . . the only way we can *act* is to endorse will to power" (207).
4. "if you have an aim of which you approve," you cannot "justifiably decline to fulfill it maximally" (208).
5. So, in acting we commit to maximal power.

For the moment let's grant Katsafanas that agential activity and power are constitutive aims of action (more on this shortly). Still, I have several

worries with this argument. First, the inference to Step 3 (on my reconstruction) assumes that endorsing an action given that it has such-and-such etiology entails endorsing the components of that etiology. This assumption is questionable, even on Nietzschean grounds. Someone who embraces Nietzsche's "*amor fati*," or his ideal of affirming life by willing its eternal recurrence, might endorse an action that was motivated by *ressentiment* while rejecting the *ressentiment* that brought it about. We need to hear more about the operative notions of "endorsement" and "approval."

Second, even granting that we aim at endorsing the etiological components of our actions, the inference to Step 3 assumes we can validly infer from 'S aims at endorsing what caused A (/the causes of A)' to 'S aims at endorsing φ', for any cause φ of A. This is highly contentious: witness the massive literatures on quantifying into attitude contexts and the logic of embedded questions. Suppose I aim to help the neediest person in the world, but I deplore showing mercy to one's enemies. If Bert, my arch enemy, is in fact the neediest person in the world, is it true that I aim to help my arch enemy? Or that I aim to help Bert? We need to hear more about the logic of aims. (Though Katsafanas often slides between 'aim'-talk and 'commitment'-talk, this is not innocent. The inference seems much more plausible when couched in terms of commitments than in terms of aims -- and much less plausible when couched in terms of, say, goals or desires.)

Third, Step 3 conflates a constitutive *aim* of action and a constitutive *feature* of action. One can aim at endorsing one's action given that it aims at power without actually endorsing the action. One's action can be unstable. It can fail to satisfy the aim of activity.

Finally, even granting that we endorse will to power, no argument is given why it is unjustifiable not to maximally fulfill aims that we endorse (Step 4). Plausibly, it would be reasonable to endorse an aim of making money without endorsing an aim of making as much money as possible.

Katsafanas adduces a secondary argument that declining to maximally fulfill the will to power is irrational: "due to the omnipresence of will to power" -- i.e., its presence in everything that we do -- "systematically neglecting its fulfillment will involve systematically ignoring reasons for action" (208). But this argument is in tension with Katsafanas's own concession that the (pro tanto) reasons generated from the will to power can be outweighed by the reasons generated by other values (197-200). (This helps him reply to the worry that we have most reason to perform any action that would generate most resistance.) If Katsafanas is right to concede this, then it needn't be irrational not to maximize power. Not performing an action that there is some reason to perform but most reason not to perform does not involve ignoring reasons.

One of Katsafanas's principal objections to Velleman and Korsgaard is that their accounts fail to generate conclusions about what we have more/less reason to do. It is thus of central importance that Katsafanas provide a concrete account of how will to power can be used to generate such normative conclusions (chs. 7-8). However, there is an internal tension in his account of how will to power assesses other values.

On Katsafanas's view, will to power has a "privileged normative status." Other values can generate reasons for action, but only insofar as they are "compatible with," or do not "conflict with" will to power (e.g., 149-152, 189-191). For a value to "conflict" with will to power is for it to be the case that "having this value entails taking there to be reason to perform certain actions which will to power entails that there is reason not to perform" (192). This way of understanding conflict is in tension with Katsafanas's claim, mentioned above, that the reasons generated by other values may override the reasons generated by will to power. It categorizes (at least some) cases where will-to-power-generated-reasons are overridden as cases of conflict. To use Katsafanas's own example (189), suppose for reductio that an agent S's value of self-preservation entails that S has (overriding) reason to refrain from sticking S's hand in a fire, whereas will

to power entails that S has reason not to refrain from sticking S's hand in the fire. By Katsafanas's definition of 'conflict', the value of self-preservation conflicts with will to power. So, it doesn't generate reasons for action, and a fortiori doesn't generate overriding reasons. Contradiction.

One way Katsafanas might revise his account would be to treat conflict in terms of a threshold and say that a value conflicts with will to power if, for a sufficient range of circumstances, it entails that we have (some, most) reason to do something that is incompatible with what will to power entails we have (some, most) reason to do. This could let Katsafanas maintain will to power's privileged normative status, while allowing that in particular circumstances it is possible to have most reason to do something incompatible with will to power -- namely, as long as such circumstances do not arise sufficiently often so that the value can retain its reason-providing force.

How Katsafanas revises his account may affect the theory's normative consequences. Katsafanas notes that "most individuals" value human life, intellectual endeavors, etc. But what about individuals who do not? Consider an agent like Susan Wolf's JoJo, who wholeheartedly endorses the sadistic values of his dictator father. Presumably, valuing murder, oppression, etc. is compatible with willing power (even if it isn't entailed by it (235)). If so, does JoJo have (most) reason to seek and overcome resistances in promoting these values? More needs to be said about how values generate reasons and the relation between will-to-power-generated-reasons and other-value-generated-reasons. Katsafanas's account thus seems vulnerable to the same sort of deficiency he bemoans in Velleman's and Korsgaard's accounts.

A central claim of Nietzschean Constitutivism is that power is a constitutive aim of action. Katsafanas is clear that power, in the relevant sense, is not brute domination. Still, it is quite a surprising claim that acting in the pursuit of an end constitutively involves aiming at seeking

and overcoming resistance to that end.

Katsafanas has two strategies for responding to apparent counterexamples. The first is to show that, contrary to initial appearances, the action does in fact aim at resistances. For example, considering the action of "watch[ing] a lowbrow sitcom on television," Katsafanas says that "there are resistances here, albeit of the most minimal sort: one must attend to the program, one must support oneself on the couch, one must resist competing desires that incline one to perform other actions, and so on" (181). Perhaps we could modify the case to stave off this sort of reply. One might be lying down with one's head perfectly angled toward the TV; one might be watching the show but be content to close one's eyes or let one's thoughts wander aimlessly should they begin to do so; and so on. But even if there are resistances involved in my action, one might worry that treating seeking and overcoming those resistances as *aims* of my action misdescribes the phenomena. Again, we need to hear more about the logic of aims.

Even if we find an action that genuinely appears not to aim at power, Katsafanas has a backup strategy: show that the action is part of a larger action that aims at power. For example, considering the action of moving a pen across a page, Katsafanas notes that this action "is typically a part of or means to some larger action, such as writing a paper," to which "many agents do aim at encountering and overcoming resistances -- such as intellectual challenges" (179). Katsafanas's reply assumes that any sub-action of an action that aims at power also aims at power. Call this the *aims transfer principle*. Consider the maximal action of living one's life. Suppose we assume, not implausibly, that every person must aim at encountering and overcoming some resistances in the course of her life. By the aims transfer principle, this would itself show that every action of every person aims at power. This is too easy a way to show that action has a constitutive aim. The aims transfer principle may trivialize the claim that action constitutively aims at power.

Though Katsafanas does think that Nietzschean Constitutivism represents Nietzsche's ethical view (243), he is explicit that his primary goal is philosophical rather than interpretive. It is sufficient for Katsafanas's purposes that his version of constitutivism be inspired by Nietzsche's texts. Nevertheless in closing I would like to briefly consider one aspect of Nietzschean Constitutivism qua Nietzsche interpretation: its ability to capture Nietzsche's claims about value creation.

There is a longstanding interpretive puzzle concerning how to reconcile three of Nietzsche's central claims about value (152):

1. Power has a privileged normative status.
2. There are no objective values, or there are no objective facts about what is valuable.
3. All values are created by human activities.

As Katsafanas pointedly asks, "If there are no objective values, and all values are created, why should power enjoy a privileged status?" (153). Katsafanas argues that we can reconcile this apparent tension in Nietzsche's ethical thought by treating Nietzsche as accepting that power is the constitutive aim of action:

power is not an objective value, in the sense that it would not have value independently of a particular feature of human activities. Rather, we are committed to valuing power merely in virtue of acting, because power is the constitutive aim of action. Thus, power has a privileged normative status. Moreover, there is a sense in which the value of power is created by human activity: the structure of our own actions commits us to valuing power. (182; cf. 163)

I am unsure whether constitutivism captures Nietzsche's claim (3) that all values are "created" (see my "Nietzschean Constructivism," forthcoming in *Inquiry*, for further discussion). As the above quote highlights, constitutivist arguments explain why we are *committed* to certain values.

They capture how "the authority of normative claims [is] to be *justified*" (1; emphasis mine). What depends on human willing, according to Nietzschean Constitutivism, is the fact that we are committed to thinking that power is valuable. This is not equivalent to saying that the value of power depends on human willing. The latter is a metaphysical claim; the former is an epistemological claim about what justifies our normative views. Nietzsche's metaphorical talk of value "creation" certainly does not wear its interpretation on its sleeve. Nevertheless more needs to be said to justify that establishing that we are committed to valuing power counts as establishing that values are "created."

This point is not merely of interpretive interest. It highlights the question of what constitutivism is an account of. The step from the claim that we are committed to the value of power to the claim that this very commitment metaphysically grounds (constitutes, etc.) the value of power is a natural one. But it is not obvious that the sorts of considerations adduced in favor of constitutivism force it upon us. Suppose we accept that power is the basic value, and that what justifies our thinking this is that power is a constitutive aim of action. Could we also treat the value of power as metaphysically grounded in its instantiating some other irreducibly normative, non-natural property? Katsafanas has made considerable philosophical progress in elucidating the proper structure of constitutivist ethical theories. But further questions regarding the scope of these theories remain.