Reflection on the self reveals a variety of puzzles: What preserves our identity through time? How do we refer to ourselves? By what means, if any, do we gain self-knowledge? But the metaphysical, semantic, and epistemological questions about the self are not exhaustive. For when one thinks, decides and acts, there is a straightforward sense in which every thought, decision and act is one’s own: one is identical, in some way that the metaphysics of personal identity will explain, with the subject of each of these. Yet not everything one thinks, decides and does is expressive of who one really is, in some other sense. Each of us has been horrified by a desire, dream or action undertaken or contemplated, and in these instances, the states and actions that are the self’s in the straightforward (metaphysical) sense are not ones that one is willing to admit, or even ought to admit, into one’s self-conception.

A moral or psychological notion of the self, distinct from the metaphysical notion, is central to the concept of autonomous agency. If this were not the case, then every uncoerced action performed by the metaphysical self would be autonomous.¹ But it is not so. We see this not only in certain instances of internal compulsion, including cases of kleptomania, obsessive-compulsive disorder, and various addictions, but also in more ordinary instances of internal distress and conflict.

The difficulty on which I focus is this: it is clear that one can be alienated from certain of one’s actions and attitudes, but it is not clear in what that alien-

¹. I use the term “coercion” broadly to refer to an extensive range of ways with which autonomous agency might be externally interfered, including the use of brainwashing, science-fiction neurological puppetry, and other forms of manipulation by outside persons and forces, such as threats, indoctrination, and post-hypnotic suggestion.
ation—or its contrary, identification with particular states—consists. Clarification of these phenomena should serve our understanding of autonomous agency. Harry Frankfurt, in particular, has through the years called our attention to this fact, claiming that the notion of identification is “fundamental to any philosophy of mind and of action.”

In this essay I defend particular accounts of the phenomena of alienation and identification, working to incorporate what is valuable about the original Frankfurtian approach into a different theory of the self. With reference to recent work on wholeheartedness and the “whole self,” I explore the issue of whether or not coherence in any sense among certain of one’s attitudes is required for autonomous action. I refine and illustrate the notion of coherence at issue.


Drawing on accounts I provide of competition among, and coherence between, preferences, I develop a particular theory of personal autonomy, one that is not purely structural, but that incorporates procedural elements.

1. INTERNALITY AND ALIENATION

Consider the following as an example of the phenomenon of alienation. Imagine that a certain personal injury attorney has brought home some work in his briefcase, including a file containing gruesome photographs of the victim of an industrial equipment malfunction. Not wanting his children to come across the photos, he hides the briefcase away in his study closet. We can suppose the attorney to be an ordinarily stable, fair, and warm-hearted man, someone who loves and cares for his family and friends, and who has cultured tastes in music, art and cuisine. He has no especially odd habits or hobbies. He is educated, empathetic, and morally sensitive.

Suppose that, as the attorney is lying in bed in the night, mentally reviewing the case, he finds himself with an odd desire: to creep downstairs to his study to gaze at the photos of the victim’s injuries. One might form such a desire, of course, out of the aim of inspiring oneself to present an emotionally moving closing argument on behalf of one’s client. But suppose this is not so. Rather, the desire of the attorney’s is only an attraction to the pleasure of peering at the photos, something he is startled to recognize. Gazing at the photos now, at home in the middle of the night, is a course of action that he takes to be neither intrinsically nor instrumentally good. The desire arrived unbidden, and he finds both it and its intentional object disgusting. To the question, “Why would I want to do that?” he can find no answer among the elements of his character.

Suppose, nonetheless, that the attorney does act on the desire. And suppose that, when his wife, having arisen in the night, finds him in the study mesmerized by the photos, his face flushed with pleasure, he is flustered, embarrassed, and at a loss to explain himself. He has now acted in a way that has violated his wife’s expectations and her trust, yet her trust is something he holds as highly valuable. His action in the night makes no sense in terms of the goals or principles he is willing to defend.

In this example, the agent does something, and no one else makes him do it. He wants to do it, and he acts intentionally. But he is divided, personally disgusted over his desire and his action. What he does is uncharacteristic of him, at odds with who others think he is and with the person he takes himself to be. The case is one of alienation from desire. The attorney is alienated from the morbid desire to peer at the photos in the night, and he does not endorse it as a reason for him to act. The desire is repulsive to the self, is not endorsed by the self, and is not something with which the self identifies.

Cases like this one illustrate a flaw in a natural initial approach to characterizing autonomous action. A natural root idea is that one acts autonomously in doing what one wants. This basic account has intuitive appeal. But autonomy cannot be the mere absence of obstacles to the fulfillment of desire because, in doing what one wants, one might frustrate one’s more ultimate ends and so, in an important sense, bind rather than liberate oneself. Since desires themselves can be enslaving, the carefree pursuit of the satisfaction of each of one’s desires is not the right model of the fully autonomous human life.

Some theorists propose to deal with this problem by offering conceptions of autonomy in terms of a “real self,” characterized on some accounts (for example, Gary Watson in “Free Agency”) in terms of values as opposed to mere desires, and on others (for example, Harry Frankfurt in “Freedom of the Will and the Concept of a Person”) in terms of higher-order, rather than lower-level, desires. Thus, using the former account, we could say that the attorney acts non-autonomously in viewing the photos in the night because he acts from an appetitive impulse rather than a value. The attorney’s alienation consists in a discord between his motivational and his valutational systems. Since his self is constituted by his values, which issue from Reason rather than Appetite, in acting contrary to his values, he acts in a way that is not self-ruled and so is non-autonomous.

Frankfurt’s original view, by contrast, describes the inner turmoil experienced by the non-autonomous agent not as a conflict between sources of motivation (Appetite versus Reason), but rather as a conflict between levels of desire. Thus, we could analyze the attorney’s case as one of action on a first-level desire for which the actor has no corresponding second-level volition (that is, a desire for a particular desire to lead one to act, when or if one acts). Since the real self on the hierarchical account is characterized by second-level volitions, the attorney does not act autonomously in viewing the photos in the night. His doing so frustrates and does not give expression to his true self.

Frankfurt’s own views on autonomy and the limits of the self have been reformulated and refined significantly in the intervening three decades since the publication of “Freedom of the Will and the Concept of a Person.” In the following section, I reflect on some of the literature this well-known article inspired, bringing out what has seemed to many to be so insightful about its approach, albeit one facing certain difficulties. In the subsequent sections, I present the alternate account of autonomous agency I aim to defend.

2. THE FRANKFURTIAN TRADITION: PROMISE AND PROBLEMS

Suppose we view ordinary desires of the first level as Frankfurt has evidently tended to view them: as mere impulses—pulls or temptations toward performing some action or other. These states might be voluntarily adopted, but normally they simply arise unbidden in response to stimuli. Frankfurt focused our attention on

4. Consider that one firmly committed to maintaining a healthful diet may yet find himself with a desire to binge on fat- and sugar-laden foods. And as Frankfurt has emphasized, an addict who hates his addiction may nonetheless repeatedly desire self-injection of a particular drug.
a striking fact about persons: that we are able to rise above such impulses, reflecting on them and forming attitudes that have as their intentional objects other (lower-level) mental states. While ordinary first-level desires may be passively acquired, and may or may not be attitudes the agent takes to count as reasons for acting or for deliberating in particular ways, second-order states seem in their very formation to involve the agent’s activity.

Perhaps this is what accounts for the initial intuitive strength of the hierarchical accounts of the self’s identity and the process of self-identification. A desire for having another desire (or for a certain desire to lead one to act, when or if one acts) is apparently not the sort of state that arrives unbidden. The circumstances of everyday life, that is, seem at first blush not to impose on us second-level desires, as circumstances might, and evidently often do, impose on us desires of the first order. For instance, if it is a hot day, I might find myself wanting to get a cool drink or wanting to go for a swim in the pool, in a natural drive to maintain stable body temperature. But ordinarily I do not “find myself” wanting to want to go to the pool. It would take some activity on my part to generate this second-order state, and so the second-order desire might seem to reflect, more deeply than the first-order desire does, who I am and what sort of person I want to be (in part, I want to be the sort of person who desires on occasion, including hot days, to go swimming).

The ability to “rise above” or to mentally “step back from” our desires and beliefs, reflecting on their character, is a noteworthy and valuable feature of our selves. It is noteworthy because it is, for all we know, unique in the natural world, and it is valuable because it is a capacity that helps us to control our behavior and so to control the direction of our lives. If one is able to get oneself to want what one wants to want, and one is able to get oneself to act as one wants to want to act, then it seems that one has achieved some control over both one’s mental life and one’s actions in the outer realm.

I have worked thus far in this section to pinpoint and elaborate the thought that there is appeal in Frankfurt’s early approach to characterizing autonomous action. But two salient problems with hierarchical accounts have received attention. As they are important in motivating the view to follow, I briefly review those problems. One is the problem of a regress of endorsing attitudes, each granting

5. Other philosophers highlight this ability. Keith Lehrer, for instance, emphasizes our possession of “metaminds,” minds capable of turning upon themselves in observation and evaluation. Lehrer, Metamind, 1990.

6. T. M. Scanlon (2002) and R. Moran (2002) read Frankfurt as suggesting that second-order states have authenticity in virtue of their being actively generated by the agent. Concerning first-order desires, the way Frankfurt puts it now is that a desire presents an agent not with a reason but with a problem, the problem of what to do with it: identify with it, throwing one’s weight behind it, so to speak, or reject it as an outlaw. He distinguishes four steps in the process of responding to this problem. See his “Reply to Michael E. Bratman” in Buss and Overton, eds., 2002, 86–89.

7. Frankfurt has recently claimed that higher-order states are not essentially active. On his current view, the alienation/identification dichotomy is not accurately characterized by the passive/active distinction. One may passively identify, he believes, for instance, in exhaustion. See Buss and Overton, eds., 2002, p. 90, note 2, and p. 225, note 1.
internality to the state at the prior level. Notice that the regress is especially problematic because ascension to higher and higher orders of desire is not only something that might occur, due to persistent self-doubt or intra-level conflict. Rather, it is something that must occur, as Frankfurt’s original account of “internality” requires it. What makes a first-level desire one’s own, and what makes it count as a reason for one to act, is that one has a positive endorsement of it, in the form of a second-level volition. But the second-level attitude itself can confer internality only if it is internal to the self; and applying the account of internality to this state requires a third-level endorsing state. What makes the third-order desire one’s own is that one has a fourth-level desire for it; and so on. Without a separate account of the internality of certain second-order desires, the regress of higher-order desires stands.

8. The regress problem is generated as, when deciding what to do, one consults one’s desires concerning what to do. But in order to avoid being wanton-like, and in the face of conflict among first-level desires, one ascends to the second level, asking oneself what one desires to desire to do. What has often been noticed is that nothing seems to prevent one’s evaluative questioning of this second-level desire in turn, so that one asks what one desires to desire to desire to do. Frankfurt’s original article observed that there is “no theoretical limit to the length of the series of desires of higher and higher orders,” and that nothing prevents a person from obsessively scrutinizing his motivations, “refusing to identify himself with any of his desires until he forms a desire of the next higher order,” nothing “except common sense and, perhaps, a saving fatigue.” Frankfurt first proposed that the series could be cut off non-arbitrarily by a person’s ‘decisive commitment’ to one of his first-level desires. But as Watson (1975), and subsequently Adrian Piper (1985), pointed out, the termination is arbitrary in the absence of any grounds for ending the ascent to higher-order desires. The hierarchical account posits no grounds, and so the regress of evaluation remains.

9. In subsequent essays, Frankfurt works to address this problem, amending his view in various ways, settling in his address “The Faintest Passion” on the idea that a psychic element is internal to the self if the agent is wholehearted with respect to it, which is characterized as a matter of the agent’s being fully satisfied with that element. Satisfaction does not require a positive endorsing state, thus avoiding a regress. Rather, Frankfurt writes: “Satisfaction is a state of the entire psychic system—a state constituted just by the absence of any tendency or inclination to alter its condition” (1999, 104). This lack of interest in making changes “must nonetheless be reflective” (1999, 105).

There is reason to doubt the success of Frankfurt’s current method for dealing with the regress problem. Specifically, certain explanations for the required “satisfaction”—exhaustion, manipulation, and frustration, for instance—seem not satisfactory authenticating states or processes for an agent’s desires. Why think that A is what I “really want” if I am only satisfied with the desire for A because I am too tired to think any harder about the matter? It seems that my autonomy would be increased by my forcing myself to contemplate further. Frankfurt believes it is irrelevant what an agent’s reasons are for joining himself to a desire, for identifying with it. But some reasons for doing so are intuitively autonomy-undermining, such as manipulation by another via hypnosis (which might occur in such a way as to preserve the reflectivity of one’s satisfaction with one’s motives), as well as guilt, threats, laziness, and inattention. It seems that one could increase one’s autonomy by combating these forces.

10. Frankfurt has been widely read as requiring positive evaluation and, indeed, given his talk of “endorsement,” “decisive commitment,” and the “desirability” of one’s desires, the reading is not implausible. In recent responses to critics, however, Frankfurt has clarified that his intention was and is not to require positive endorsement for identification. Rather, he favors an account of identification as acceptance. See Buss and Overton, eds., Contours of Agency, 2002, especially pp. 160–61.
This regress of practical reasons is interestingly parallel to the regress of theoretical reasons in the epistemic realm. In response to questioning of the status of some desire to count as authentic or end-setting for action and practical deliberation, one cites a desire in favor of that desire. Likewise, in response to skeptical questioning of the justification of some belief, on an internalist conception of justification, one cites a reason in favor of that belief. Both evaluative projects, practical and theoretical, generate a problematic regress of reasons. Below I return to this parallel, as reflecting on it turns out to be illuminating.

A second difficulty for the hierarchical approach may be termed the problem of arbitrariness. Consider Frankfurt’s “willing addict,” one who both desires to take a particular drug and desires that this desire be effective in leading him to act. In saying that the willing addict acts freely because in taking the drug, his will is “his own,” Frankfurt implies that the real self is to be identified with one’s second-order volitions. But what makes this a plausible, non-arbitrary conception of the most intimate self? Second-order desires are, we notice, just desires, and Frankfurt poses no requirements on their formation. He does not, for example, require that they be formed carefully or with sober judgment or in calm and reflective circumstances. Second-order volitions, on Frankfurt’s view, do not necessarily represent the agent’s highest ideals, values, or standards. The agent may offer any poor reasons in their support and, in fact, may have no defense at all available in their favor.

Furthermore, as Watson and others have emphasized, second-level desires could be imposed by a neurosurgeon or a demon or a hypnotist. Nothing about their *level*—the fact that they are desires having as their intentional objects other desires or states of affairs concerning desires—precludes this. A hierarchical account, then, appears to require an added non-coercion (or external manipulation) condition. Frankfurt disagrees; his most recent work reaffirms his commitment to a purely structural account of autonomy, one with no historical requirements on the relevant mental states. But this is counterintuitive, for an imposed character—even one exhibiting a high degree of internal harmony—seems an improper candidate as a ground for genuinely autonomous action. In sum, second-order volitions are simply desires with a particular sort of intentional...

11. As Watson argues, higher-order desires are just another species of desire, and “nothing about their level gives them any special authority with respect to externality. If they have that authority they are given it by something else.” G. Watson, 1987, p. 149.

12. In a striking passage, Frankfurt writes: “It may not be from the point of view of morality that the person evaluates his first-order desires. Moreover, a person may be capricious and irresponsible in forming his second-order volitions and give no serious consideration to what is at stake . . . There is no essential restriction on the kind of basis, if any, upon which [second-order volitions] are formed.” H. Frankfurt 1971, in Frankfurt 1988, p. 19, note 6.


14. For instance, Frankfurt writes, “it seems to me that if someone does something because he wants to do it, and if he has no reservations about that desire but is wholeheartedly behind it, then—so far as his moral responsibility for doing it is concerned—it really does not matter how he got that way.” “Reply to John Martin Fischer,” in Buss and Overton, eds., 2002, 27.

object, and arbitrariness seems to infect the claim that they constitute the agent’s authentic self.  

One line of response to the arbitrariness problem would be to locate the real self elsewhere: some free will theorists, for instance, understand the term “decision” as the active formation of intention to act, viewing states of intention (rather than desires or judgments or volitions or something else) as the outputs of all decisions; and the self is implicitly understood to be that which is active in the formation of decisions to act.  

But notice that there is a problem of arbitrariness for accounts centering on the notion of decision, as well: Why think that all decisions (so conceived) are internal to the self, that the class of decisions does not allow for division between internality and externality?  

Certainly there may be decisions from which one is alienated: for instance, the decision to lash out with angry words at one’s colleague, which one later and, even at the moment of action, very much regrets. And consider the addict who hates his addiction. Suppose that he is horrified by his drug-taking and by his desire to take the drug to the extent that he regards his drug-taking behavior as expressive of his addiction and not of who he really is. Nonetheless he might decide to re-inject heroin. This decision is one that is not expressive of the actor’s self. This case and others show that the notion of identification is not easily reducible to the notion of decision to act.

The problems of regress and arbitrariness concerning identification are rather interesting problems, and every autonomy theorist must have some sort of

18. The view that decisions adequately settle the matter of identification is a view once taken by Frankfurt (subsequent to his “Freedom of the Will” paper) and later rejected. In his “Identification and Externality” (1977), Frankfurt suggests that “decisions, unlike desires or attitudes, do not seem to be susceptible both to internality and to externality” (reprinted in Frankfurt 1987, 68). Later, in his “Identification and Wholeheartedness” (Frankfurt 1987), Frankfurt proposes an account of identification in terms of decision, remarking that “it is characteristically by a decision . . . that a sequence of desires or preferences of increasingly higher orders is terminated . . . The decision determines what the person really wants by making the desire upon which he decided fully his own” (reprinted in Frankfurt 1987, 170). In “Identification, Decision, and Treating as a Reason,” Michael Bratman works to improve upon Frankfurt’s suggestion in these papers. Bratman proposes that to identify with one’s desire is to decide to treat that desire as reason-giving in one’s practical reasoning, planning, and action, to be satisfied with that decision, and either to treat that desire as reason-giving or, at least, to be fully prepared to treat it as reason-giving should a relevant occasion arise.
19. In his “Hierarchy, Circularity, and Double Reduction,” Bratman suggests that agential identification with a desire consists in the agent’s having a non-instrumental higher-order self-governing policy, with which she is satisfied, in support of that desire’s functioning, by way of that very policy; as end-setting for practical reasoning (Bratman 2002, 77). Bratman understands “policies” as intentions that are appropriately general (p. 83, n. 34, 2002). Self-governing policies have *authority* for grounding the agent’s point of view (or establishing internality to self), provided they are policies with which the agent is “satisfied” (i.e., free from significant challenges from other relevant higher-order policies). Bratman proposes, in virtue of the fact that they play a central role “in the constitution and support of the psychological continuities and connections highlighted by broadly Lockean approaches to the agent’s identity over time” (p. 76, 2002). The policies, for instance, help to organize an agent’s action over time.
response to them. For autonomy is, at base, self-rule, self-government or self-determination, as opposed to enslavement or rule-from-without. As such, it requires for elucidation some account of what is internal and what is external to the self. Frankfurt gave us a split-level approach to drawing the most difficult aspect of this distinction. Should we now abandon or embrace the approach of the Frankfurtian tradition for characterizing autonomous agency?

3. PREFERENCE AND IDENTITY

Perhaps we can retain the insights of Frankfurt’s initial approach to autonomy, yet refine it in ways that circumvent the problems of arbitrariness and regress. What is needed in the face of these problems, I believe, is defense of a conception of the self which incorporates our higher-order reflective abilities, yet which is more expansive than Frankfurt’s previous view of the self as delimited by desire. In particular, as Watson once emphasized, the central role that evaluative reasoning plays in self-construction should take center stage. Expanding on the earlier views of both theorists yields a view on which the self is taken centrally to include not only certain desires, but also other important types of attitude and ability. The account of autonomous agency I wish to defend takes autonomous agents to be moved by desires, rather than by judgments, and thus may be described as broadly Humean. But the account is non-Humean and non-Frankfurtian in requiring a standard of evaluation against which desires are assessed, giving a central and substantive role to Reason or the understanding. Reason not only establishes methods of satisfaction and settles conflicts of priority among authentic desires, but also grants the stamp of authenticity to certain ones and not others by making judgments of worth.

Suppose we use the term “preference” as a term of art, letting it stand for a desire formed by a process of critical evaluation, in particular, with regard to the


As noted above, if Frankfurt ever did require reflective endorsement for identification, he no longer does. Frankfurt’s most recent work makes quite clear his rejection of a ‘positive evaluation’ reading of identification in favor of identification as acceptance. Thus I am quite sure that Frankfurt himself would reject the positive evaluation of, and amendments to, the “Frankfurtian tradition” that follow.

21. The limits of the self are not desires, on Frankfurt’s most recent view; the self’s limits are defined by its volitional necessities, characterized by a complex, double inability to will certain states of affairs.

22. Frankfurt would likely call this analysis part of an overly intellectualized picture of the mental life. See his “Reply to Michael E. Bratman” (Frankfurt 2002, 89).
agent’s conception of the good. A preference is an evaluated desire, one actively generated in light of the agent’s beliefs and, especially, as part of his pursuit of the good. Such reflective desires have as their intentional objects, sometimes, a particular course of action and, other times, the state of affairs of a certain first-level desire’s being satisfied in action. Unlike Frankfurt’s second-order volitions, preferences are identified as such by their mode of generation and not by the type of their intentional object.

The deliberative, evaluative process preceding preference formation need not, although it might, be protracted, and it need not, although it might, take place at the fully conscious level. Just as one might believe some proposition for the sake of the argument or out of fear, and not because one thinks it is true, so one might desire something out of guilt or instinct or curiosity, and not because one thinks it is good. Morbid desires, as in the case of the attorney above, are ready examples. If one, like the attorney, desires to peer at a bloody scene, yet this desire is not formed through critical evaluation with respect to worth—and, in fact, upon reflection one finds neither that course of action nor that desire to be either instrumentally or intrinsically good—then the desire to peer is not a preference in the stipulated sense. David Gauthier similarly uses the notion of a considered desire to mark a difference between desires that impel a being to act by sheer force of their strength, on the one hand, and desires, on the other hand, that motivate to action only after having survived a process of reflection in light of one’s beliefs, including beliefs about the nature and quality of various experiences.23 Such reflection sometimes alters what one desires. For instance, one might desire to run for election to a political office; but upon reflection on what it is like to live in the public eye, one’s life subjected to near-constant scrutiny, one might form a considered desire not to run.

On my view, these attitudes of preference are central to autonomous agency. This is the case, I argue, because it is reasonable to conceive of the self as constituted by an aggregate of preference and acceptance states, along with a certain capacity.24 Suppose we use the term “acceptance” to mark the mental endorsement of a proposition formed by critical reflection with the aim of assenting to what is true.25 Then beliefs form the wider class, and acceptances are a subset of these. The proposed account of the self provides a more expansive notion than that implicit in Frankfurt’s original theory, and it is one that need not bear any special ontological commitments for selfhood.26

To elaborate: through time, each of us refines our understanding of the world. We adopt new convictions, and we reject or retain others. We also form new tastes, reformulating our tastes through time, coming to prefer certain activities and ways of life. All the while we shape the kinds of persons we are, whether bul-

26. The view does not commit us, for instance, to the existence of an agential power to cause events in such a way that is not reducible to purely event-causal terms, nor does it require commitment to a trans-empirical Self or a Cartesian mind.
lying, kind, cold, or adventurous. Our ability to form evaluative states of mind is essential to being someone. The faculty is critically evaluative: it enables us to subject our motivations to scrutiny in light of what we accept, and so to come to view certain attitudes as genuine reasons for acting and deliberating in certain ways. It is this faculty for evaluative reasoning that I view, together with an aggregate of preferences and acceptances, as constituting a person’s moral or psychological identity. Since the relevant attitudes represent what a person reflectively wants and what he believes when aiming at truth, we might call the collection of preferences and acceptances of a person’s psychology his character. The attitudes constitute what he is like. Then, we can understand the self as a character, along with the capacity or faculty for forming and reforming that character. 27

What makes plausible this conception of the self? First, although it remains controversial whether our faculty for forming considered states of mind is the ultimate determinant of the contents of our characters—in a sense implying the lack of constant causal necessitation by the past and the natural laws—nonetheless it is appropriate to consider a faculty for shaping the elements of the character as central to oneself. Our moral and psychological identities are connected to our ideals, but not only these. Fundamental to our having ideals at all is a capacity for forming conceptions of what is right, true and good and for evaluating states of affairs, including states of mind and courses of action, in light of those conceptions. Whatever the deliverances of our backgrounds, we have an ability to evaluate desires with respect to worth and to critically evaluate propositions, granting some our acceptance. As we make such decisions from the perspective of our existing characters, we reform ourselves, becoming to a greater extent self-made.

Second, unlike the hierarchical account, the proposed account takes our selves to be structured not solely by desires, assuming that our convictions concerning the truth partially define us. I might accept that Democratic candidates are generally superior to Republican ones. You might accept that the Washington Redskins are the superior football team, while another accepts that being an ardent fan of any professional athletic team is a spectacular waste of time. Such attitudes feature in the characterization of who one is—they help to pick out the individual who is oneself.

Further, our characters include attitudes formed as we reflect on what is true and good, and not those that we adopt out of unevaluated whim, fashion, fear, guilt, instinct, or convenience. It is fitting that the elements of the character are not passions that overtake one or beliefs haphazardly adopted, but rather are beliefs and desires endorsed by one’s own evaluative faculty, since a character is the complex of attributes or features that mark and distinguish an individual from others, and states formed in light of our subjective conceptions of truth and goodness both exhibit variation from others and maintain some stability through time.

27. Eleonore Stump proposes a similar revision of Frankfurt’s account (1988). Building on Aquinas’ idea that an agent wills to do some action \( p \) (or bring about some state of affairs \( q \)) only if the agent’s intellect at the time of the action represents \( p \) (or \( q \)), under some description, as the good to be pursued, Stump revises Frankfurt’s theory to take account of the role of intellect. The proposal I develop here fills in a theory of the self and does not rely on the existence of “the will,” understood as a hunger or drive for the good.
The proposed view takes centrally into account the role that evaluative reasoning plays in our practical deliberations, strengthening the Frankfurtian account by giving the evaluative standard according to which we form desires, including those concerning which desires lead us to action. If values can be understood either as preferences in the stipulated sense or as acceptance states concerning what is good, then the account takes values centrally into the account of the character. The view, however, has us structured neither solely by desires, nor solely by values, but includes our perspective on the truth about all sorts of matters and encompasses our faculty for reflective evaluation.

The account of the self partially in terms of preferences can be put to work in understanding problematic cases of alienation, as in the case of the attorney above. When the attorney acts on the desire to peer at the bloody photos in the night, he does not act from a preference, but rather from an unendorsed impulse. Constitutive elements of his character include the preference to remain in bed, the acceptance that gazing at the injuries now is not something useful or worthwhile to do, and the preference for maintaining both his dignity and his wife’s trust. Hence the morbid desire is one from which he is alienated, in that it is not part of his psychological identity. Acting on it frustrates the self.

For several reasons, then, I take the proposed account of the self to be a non-arbitrary one. Notice that we need not view the self on this conception as an “entity,” implying anything in particular for our ontology. The self as conceived is, rather, a kind of psychological essence. We often make remarks indicating our commitment to such motivational essences. We say concerning particular individuals, for instance, “he is no longer himself,” or “she is not the person she used to be.” In so doing, we make no unusual metaphysical claims. We use such locutions, rather, to mark significant changes in moral or psychological identity, as in instances in which a person who was once malicious and calculating undergoes a personal transformation, becoming kind and generous.

It may be that the proposed account of the self is yet insufficiently expansive. It may, for instance, problematically leave aside emotions. There is significant controversy over what constitutes an emotion, and I leave the issue aside here. In addition, Michael Bratman has, in a series of recent works, emphasized the need for the structures of plans and policies in an adequate theory of agency. If our plans and policies are central to who we are, and these cannot adequately be understood in terms of the notions of preference and acceptance, then the pro-

28. For one recent approach to characterizing the nature of values, see Michael Bratman, “Valuing and the Will,” Philosophical Perspectives 14 (2000): 249–65. Bratman suggests roughly that, for one important kind of valuing, an agent values X when he has a desire for X and a self-governing policy in favor of treating that desire as providing an end that is justifying in motivationally effective deliberation (Bratman 2000, 260).


posed conception of the self requires augmentation with additional mental structures.

It is a short step from the proposed account of the self to a theory of autonomy. For if we consider autonomy to be self-direction or self-rule, then relying on the account yields an understanding of autonomous action as follows: Autonomous action is action on an uncoercively formed preference. Since the capacity for critical reflection is partially constitutive of one’s moral or psychological identity, one’s exercises of this capacity are “one’s own,” barring external manipulation via coercive mechanisms. This latter concern motivates the non-coercion condition. To adopt the proposed approach is to take an evaluative—some might say “rationalistic”—approach to autonomy. Yet surely it is neither surprising nor untraditional to view Reason or the Understanding as establishing a person’s “take” on the world, his perspective, or his mode of making sense of his experiences. Correspondingly, behavior is appropriately viewed as animalistic when not preceded by reflective endorsement with reference to a person’s understanding of the good.

32. It may be that the notion of a standing preference (for instance, to perform act A not at present, but when circumstances X arise) and the acceptance of certain procedural and future-oriented principles, such as a principle of avoidance-of-later-regret, go some way toward addressing Bratman’s concerns. I use the notion of a standing preference in response to certain objections to a particular account of free will in Ekstrom 2000.

33. Some theorists may argue for an additional, metaphysical requirement on autonomy, namely, that the uncoercively formed preference preceding autonomous action itself be indeterministically caused by certain events. They may urge this requirement on the basis of the thought that autonomy—understood as roughly equivalent to freedom of the will—requires the falsity of the thesis of causal determinism: the thesis that there is, at every moment, exactly one physically possible future. See, for instance, Susan Wolf’s treatment of what she terms “the Autonomy View” in her 1990; see also Ekstrom 2000, chapter 4, in which I understand free action as action non-deviantly caused by a preference with undefeated authorization, which requires that a preference be uncoercively formed and indeterministically caused by considerations of the agent’s.

Delineating the relation between autonomy and free will is made complicated, of course, by diverse uses of the terms. Nonetheless it is clear that autonomy theorists and free will theorists are at base commonly interested in the power of agents to control, or to be in charge of, their deliberations, decisions and other actions. In ordinary contexts, ascriptions of autonomy are not sensitive to the truth or falsity of metaphysical doctrines such as the thesis of causal determinism and the doctrines of divine providence or divine foreknowledge—though, of course, they might become so, if pressed. Consider medical and legal contexts, for example, in which we are concerned that a patient give his autonomous, informed consent to undergoing a certain procedure, or in which we aim to ensure that an accused individual’s confession is autonomously given. We are usually in such circumstances not preoccupied with metaphysical questions concerning the workings of the universe, but rather with promoting an agent’s ability to decide on her own, rather than on the basis of someone else’s authority, and not from some motivation that is internal in a broad sense but not an important narrow sense. Perhaps one type of free agency—the type grounding deep responsibility understood in terms of desert—does require the falsity of causal determinism. (For defense of this view see Ekstrom, 2000, chapters 5 and 6; and Ekstrom, “Libertarianism and Frankfurt-Style Cases,” in Robert Kane, ed., The Oxford Handbook of Free Will [Oxford: Oxford University Press, 2002], pp. 309–22.) But in wide uses of the term ‘autonomous,’ one may act autonomously regardless of the truth or falsity of this doctrine.
4. COHERENCE AND PERSONAL INTEGRATION

I have thus far suggested a move from desire to preference in conceiving of the self and have defended the resulting account as non-arbitrary. Is this marriage of Reason and Appetite sufficient for characterizing agent autonomy? Or are there further conditions, perhaps structural conditions, concerning the relations among preferences themselves? Does a preference in itself have insufficient authority to stand for the perspective of the agent—to indicate what “she really wants” or to count as the desire with which she identifies?

Recall the evaluative regress problem for hierarchical accounts of autonomy, parallel to the epistemic regress of reasons in favor of belief. In response to questioning or doubt over the authenticity of some desire, the hierarchical account posits an ascension to higher and higher levels of endorsing desires. But a coherence approach to autonomy can take its cue from the coherentist response to the epistemic regress, providing a parallel account of the authorization of preference. When we wonder about whether some preference is worthy, whether or not it is one we ought to have, we need not ascend to higher levels of desires for desires for desires. Rather, we can consider the preference to be a valuable one to have just in case our other preferences and our acceptances can be used in its defense—that is, if they support our adopting or maintaining it. And we might see the preference as supported by our other preferences and by our acceptances just in case those states form a network of support, a coherent and interwoven structure.

Suppose that we take an agent’s true or most central self to be the capacity for forming and reforming the character, along with a subset of the acceptances and preferences, namely, those that cohere together. One’s preferences, we might say, are personally authorized—or sanctioned as one’s own—when they cohere with one’s other preferences and acceptances. The incoherent preferences we might see as elements of the self, but of the peripheral rather than the central self. If autonomous action is action directed only by the true or central self, then autonomous action requires action on an uncoerced and cohering preference. In the remainder of this section, I investigate the question of whether it is sensible to maintain any type of coherence requirement for autonomous action.

Certainly there are legitimate questions concerning whether or not we have good reason to posit such an entity as a true, or most central, self. Once we have located the self, it may seem grandiose to suppose that there is some sort of split within the agent. Does the existence of a true self imply the existence of a false self? If so, what is a false self? And if there is such an entity as a central self, then is there also a peripheral self? To what use could we put the notions of a false or a peripheral self?

Notice, however, that it is often reasonable and appropriate to say that some elements of person’s psychology are more central, while others are peripheral. The value a parent places on the health and well-being of her children, for example, might be utterly central to her identity, whereas the acceptance that it will rain tomorrow might be peripheral. Setting on the right way to account for this difference in depth or centrality would be useful. Let me elaborate a bit on considerations favoring some sort of structural requirement on the elements of the
central self and, in particular, on considerations favoring a coherence picture of that structure.

One might defend the notion of the central self as a subset of the self—the deliberative faculty plus the cohering preferences and acceptances—on several grounds. First, since cohering elements hold together firmly, displaying consistency and mutual support, they are particularly long-lasting. Acceptance states that are part of a network of other acceptances and preferences with which they fit tend to remain relatively immune to change—perhaps they are not constant, but they are at least unlikely to be in a state of perpetual fluctuation. Coherent preferences, likewise, are guides to action that will likely remain, as they are well supported by reasons in their favor. This seems intuitively correct. The core of one’s character is not permanently fixed, yet it should be relatively stable through time.

Second, it is the cohering elements of the character that are not directly opposed to each other but instead have some support, so that they are fully defensible. Sometimes, to be sure, first-level desires (such as an unwilling addict’s desire to take the drug) are quite long-lasting, and painfully so. Nonetheless, since such desires do not cohere with the character, they are not fully defensible on the basis of it. Elements of the character should be beliefs and desires of ours for which we can provide defense in the face of external challenge and neurotic doubt. And it is only the cohering elements of the character, not the anomalous parts of it, for which we can mount a wholehearted defense. For example, one might have nothing at all to say in favor of one’s desire to buy a lottery ticket, as one might have all sorts of beliefs and desires that oppose it. Were one to evaluate positively the desire in this instance, forming a preference to buy the ticket, still that preference might fail to cohere, because one might have, as well, a preference to spend one’s money wisely, another preference to take few risks, and an acceptance that the odds of getting a winning ticket are extremely low.

One might argue that one could still act freely in buying the lottery ticket, even though the preference for doing so is not a cohering one. Perhaps this is correct concerning free action. Yet autonomy may be a thicker notion, and the argument for the non-autonomy of the act of buying the ticket does have some force. Suppose one acts on a non-cohering preference for buying the ticket. Then buying the ticket is something not only unusual but also perplexingly odd for one to do: It does not fit with the bulk of one’s evaluated reasons. An observer of one’s ticket purchase might remark, “How strange, that is not like him at all. Not only have I never seen him buying a lottery ticket, but in fact he has told me that he opposes the state lottery, that he prefers to be careful with his money and to spend extra cash on small gifts for his children. I wonder what has come over him.” If the preference on which one acts is an incoherent one, then one is divided in performing the act; one suffers internal conflict over what one does. It may be that this level of conflict is sufficient to undermine autonomy.

35. From my perspective, this would require the absence of coercion and the falsity of causal determinism, as well.
Third, not all elements of the character, but only the coherent aspects of it, are features that one is comfortable owning. One tends not to have distress over attitudes that cohere with one’s other considered attitudes, as one might have distress or consternation over incoherent or anomalous attitudes. In acting on a cohering preference, one acts without significant mental reservation, and one is wholeheartedly behind what one does, in a structural sense. Since the coherent elements fit with the other items one accepts and prefers, in acting on them, one is not frustrated or in tumult. This lack of conflict generates a state of inner tranquility, a kind of liberating comfort with oneself.

Of course, one will object that we are all conflicted to greater or lesser extents, and that we simply live with these conflicts and are able to act autonomously in spite of them. Notice, however, that I have not implausibly claimed that a self ceases to be a self if there are any conflicts among its attitudes. I have suggested that we can consider the self to be a certain deliberative faculty along with all of a person’s acceptances and preferences. So one might both prefer a particular course of action \(a\) and prefer an opposing course of action \(b\) at the same time, where both preferences count as aspects of who one is. Nonetheless, I am pursuing the idea that the core of a self might be taken to be preferences and acceptances that cohere together. When one both prefers \(a\) and prefers \(b\), and when one of these preferences coheres with the character but not the other, the cohering preference is central to the self, and the incoherent preference is peripheral.

To be sure, by adding a coherence requirement, we get quite a stringent account of autonomy. Nonetheless it may be a useful account. Consider a medical context in which one is charged with the task of making a treatment decision on behalf of a loved one who is incapacitated. One might seek to discern what the other would autonomously decide, were she able to decide for herself and were she able to let one know of her decision. Of course, so far as possible, one would take into consideration her acceptances concerning the truth, as well as her preferences. But what if one knew her to be conflicted, to have preferences and acceptances pointing in different directions? In response to this problem, one might to the best of one’s ability attempt to discern a core or bulk of her preferences and acceptances concerning features relevant to the situation, those that fit or hung together. The relevant cohering attitudes, one might think, are more at the core of her personality. They support each other. They are states by which she makes sense of the world and by which she, for the most part, guides her behavior. They constitute parts of her coherent outlook, the most important aspects of who she is. One might, then, make a decision as best one can, in light of these cohering attitudes.

I have so far suggested an intuitive understanding of coherence as a matter of mental states’fitting together, clinging to or supporting each other. Other autonomy theorists have made similar appeals to an ideal of agential wholeheartedness or personal integration.36 Frankfurt, for instance, has highlighted the notion of

36. Harry Frankfurt, “Identification and Wholeheartedness”; and Nomy Arpaly and Timothy Schroeder, “Praise, Blame, and the Whole Self.” J. David Velleman argues that what tempts us to an integration condition on psychological states is merely a perverse wish to be unconflicted. He calls an account of autonomy requiring wholeheartedness a “defensive fantasy” (p. 109, 2002)
wholeheartedness, suggesting that the wholehearted agent is free of ambivalence, such that at least one of a pair of conflicting psychic elements is rejected “as an outlaw.” How might we gain a more fully developed understanding of wholeheartedness?

It is striking that those working on these issues in moral psychology and action theory have not made use of the rich conceptual resources of epistemology, in particular, detailed accounts of coherence among mental states. In refining the notion of coherence that may be relevant to autonomy, we might, for instance, rely on the account presented in Laurence Bonjour’s *The Structure of Empirical Knowledge*. On that account, several conditions increase or decrease the coherence of a belief system, such as the number and strength of inferential connections between the system’s components, the presence of unexplained anomalies in the content of the system, and the extent that it is divided into subsystems of unconnected beliefs. Bonjour maintains that a system is coherent only if it is logically consistent, and in proportion to its degree of probabilistic consistency. Coherence, he says, is a matter of how well the various components of a system “agree or dovetail with each other, so as to produce an organized, tightly structured system . . . rather than either a helter-skelter collection or a set of conflicting systems.”

An alternative approach is the coherence notion in Keith Lehrer’s theory of knowledge. Lehrer envisions epistemic justification as a game or contest with a skeptical interlocutor who questions one’s states of acceptance. Acceptances cohere with one’s background system when they can be defended against the skeptic’s challenges by citing elements of the system on their behalf. An acceptance state coheres with a system of acceptances just in case competing propositions are, in his terms, beaten or neutralized. This account fits naturally with the way I have described the coherence of the elements of the character: as a matter of mutual defense in the face of external challenge or internal doubt.

5. THE COHERENCE ACCOUNT OF AUTONOMY

Recall that I have described a character as a certain system or amalgam of preference and acceptance states. Taking a cue from Lehrer’s account, here, then, is one way one might proceed in more precisely characterizing preferential coherence with a character. Rather than resting on a basic unanalyzed notion of *reasonableness*, as is featured in the account of epistemic justification, I take as a base notion that a preference might be more or less *valuable* to have on the basis of a character at a time. A preference is valuable to have according to a character system if it makes sense to have that preference based on that system or, in other words, if the preference is useful or worthy, according to that system at a particular time.

Coherence with a character may be formally defined as follows:

---

and a case of “wishful thinking” (p. 99, 2002). (These comments are directed against Frankfurt.) Frankfurt replies in 2002, pp. 124–28.

coherence: a preference for $x$ coheres with the character system of $S$ at $t$ if and only if, for any competing preference for $y$, it is either (i) more valuable for $S$ to prefer $x$ than to prefer $y$ on the basis of the character of $S$ at $t$, or (ii) as valuable for $S$ to prefer the conjunction of $y$ and a neutralizing state $z$ as it is for $S$ to prefer $y$ alone on the basis of the character of $S$ at $t$.

A character system can yield the result that one preference is more valuable than another by informing a person concerning what else she prefers and concerning what she accepts about the world. Her acceptances include those regarding the nature of various experiences, what it is good to pursue, and the methods likely to work in achieving satisfaction of her desires. Suppose that system $C$ is an agent $S$’s character system. Then one preference competes with another for $S$ just in case the following definition is satisfied:

competition: the preference for $y$ competes with the preference for $x$ for $S$ on system $C$ at $t$ if and only if it is less valuable for $S$ to prefer $x$ on the assumption that $y$ is good than on the assumption that $y$ is bad on the basis of $C$ at $t$.

One can meet the challenge of a competing preference either by defeating it or neutralizing it. Defeating and neutralizing are defined as follows:

defeating: the preference for $x$ defeats the preference for $y$ for $S$ on $C$ at $t$ if and only if the preference for $y$ competes with the preference for $x$ for $S$ at $t$, and it is more valuable for $S$ to prefer $x$ than to prefer $y$ on $C$ at $t$.

neutralizing: $z$ neutralizes the preference for $y$ as a competitor of the preference for $x$ for $S$ on $C$ at $t$ if and only if the preference for $y$ competes with the preference for $x$ for $S$ on $C$ at $t$, but the preference for the conjunction of $y$ and $z$ does not compete with the preference for $x$ for $S$ on $C$ at $t$, and it is as valuable for $S$ to prefer the conjunction of $y$ and $z$ as it is to prefer $y$ alone on $C$ at $t$.

To illustrate these accounts of the competition of preferences and the defeat and neutralization of competitors to cohering preferences: Suppose that one is deciding where to spend one’s sabbatical year. Naturally, a desire arises to stay put, to live at home and to work at one’s home institution, as this is the easiest thing to do. But one also has a romantic vision of spending the year, instead, in a small village in southern France. Suppose that one reflectively evaluates each of these desires with respect to one’s conception of the good, and each of them survives, such that one now has a preference for staying home (there are certain considerations showing this to be good), as well as a preference for spending the year in France (there are other considerations showing this to be good). The preference for spending the sabbatical year in France competes with the preference for spending the year at home, since it is less valuable to prefer staying home on the assumption that spending the year in France is good than on the assumption that spending the year in France is bad, on the basis of one’s system of acceptances and preferences. (That is, it makes less sense or is less worthwhile to prefer staying home, if
one assumes spending that same year in France to be good, than if one assumes
spending that same year in France to be bad.)

Now in the face of these competing preferences, one might check to see
which better coheres with what else one accepts and prefers. Suppose that one
enjoys travel. But one also has young children, and one accepts that traveling with
young children is quite difficult. In fact, uprooting them—taking children away
from their own bedrooms, toys, friends, routines, and classes—is disruptive to the
functioning of the entire family. Doing so in the past has resulted in problem
behaviors and illness. Suppose one accepts that one does not work well when ill
or in conditions of stress; that one accepts that the well-being of one’s family
members is crucial to one’s ability to focus on one’s intellectual projects; and that
one accepts that, however beautiful life in southern France may sound, traveling
to France for the year will lose one at least a month of work time to the practical
necessities involved in making the move there and back. One prefers not to lose
time in getting engaged in one’s academic projects. One prefers maintaining family
health and stability, even at the cost of some adventure. In such a case, one’s pref-
erence for staying home defeats the preference for spending the sabbatical year
in France. The preference for staying home better coheres with one’s character
system; it is supported by the elements of that system, and the competing prefer-
ence is not.

Consider a different case of competing preferences. The preference for main-
taining one’s health (preference $y$) competes with the preference for purchasing a
large, buttery frosted case from the bakery (preference $x$), since it is less valuable
for one to have preference $x$ on the assumption that $y$ is good than on the assump-
tion that $y$ is bad. Preference $y$ might not be defeated on the basis of one’s charac-
ter system (perhaps one does not have a sufficient case to mount from one’s
acceptances and preferences in favor of the value of making the bakery purchase
over the value of health). But the competitor might be neutralized by the desire to
share the purchased cake with several of one’s friends (call this desire $z$). Desire $z$
neutralizes $y$ as a competitor to $x$, since the preference for maintaining one’s health
competes with the preference for purchasing the buttery cake, but the preference
for the conjunction of $y$ and $z$—maintaining one’s health and sharing the purchased
cake with one’s friends—does not compete with $x$; and it is as valuable to prefer
maintaining one’s health and sharing the purchased cake with one’s friends, as it is
to prefer maintaining one’s health alone. Desire $z$, then, gives one a way to have
one’s cake and eat it, too. The competing preference thus neutralized, the prefer-
ence for purchasing the cake remains coherent with one’s character.

When considering what preference one ought to adopt, one takes into
account the elements of one’s character system, checking to see which of the com-
petitors it is more valuable for one to prefer. The competitors are compared against
a system of what one prefers and accepts. The preference is personally authorized
when competitors to it are defeated or neutralized.

The account of personal authorization, then, is as follows:

*personal authorization:* $S$ is personally authorized at $t$ in preferring $x$ if and only if the preference for $x$ coheres with the character of $S$ at $t$. 
Equivalently:

**personal authorization**: S is personally authorized at \( t \) in preferring \( x \) if and only if every preference that competes with the preference for \( x \) for S on the basis of the character of S at \( t \) is defeated or neutralized on the basis of the character of S at \( t \).

We may now give a coherence theory of autonomous action in terms of these definitions. On the coherence account, one acts autonomously when one’s act is non-deviantly caused by an uncoerced, personally authorized preference. A preference that is personally authorized for one counts as truly one’s own, as one that one really wants to have, since it coheres with what else one prefers and accepts. Thus action on such a preference is self-directed or self-ruled, rather than heteronomous.38

The notion of **identification** may be understood on the coherence account as follows. To identify oneself with some desire is to have a personally authorized preference for that particular desire to be the one that leads one to act, when or if one acts. To identify oneself with some belief is to have an acceptance regarding the content of the belief that is coherent with one’s character system. And to identify oneself with some course of action is to perform that act because one has a personally authorized preference in its favor.

Notice the following concerning the view. First, it is not content-specific, in that it does not require an autonomous agent to perform only certain types of actions or to form only certain types of preferences. Some autonomy theorists would like to call the subservient housewife, for instance, a non-autonomous agent, even if she has been uncoerced in adopting her way of life and even if she has chosen it as part of her conception of the good. Autonomous agents must not make self-limiting choices, they say; one cannot autonomously be in an oppressive situation. But requiring others’ choices to conform to our own conceptions of what is good, liberating or fulfilling incorrectly limits the scope of autonomy. In serving her husband and children, the housewife acts autonomously provided that she acts on a preference for doing so that was uncoercively formed and that coheres with her character. For her, serving her husband and children is a part of the pursuit of the good; from her point of view, there is nothing wrong, and in fact there is good, in serving others.

Second, the account is not a purely structural one. Entitling it “a coherence theory” should not obscure the fact that the account poses requirements on the causal history of autonomous actions: The preferences preceding autonomous actions must not be coercively formed; and procedural elements are built into the notion of preference itself, since preference states to count as such must be formed by a process of critical reflection with regard to the agent’s conception of the good.

---

38. Notice that the definitions of coherence and personal authorization leave room for outlying preferences—those largely unrelated by content to many of the other elements of one’s character—and for small clusters of related and mutually supporting preferences to count as elements of the true self, given that such preferences may lack competitors.
In fact, any purely structural account of autonomy will be unsuccessful, since the history or source of one’s motivational states is relevant.

6. INTEGRATION AND PSYCHOLOGICAL DEPTH

Consider a related view in the recent autonomy literature, “the Whole Self theory” proposed by Nomy Arpaly and Timothy Schroeder. As this view rests on a notion of psychological integration, understood in a somewhat similar way to the coherence among character elements I have addressed, comparison may be fruitful. Arpaly and Schroeder maintain that one’s meriting praise or blame for an action depends largely on the extent to which one’s self is expressed in one’s action, and they hold that the extent to which an action expresses the agent’s self is not determined by the particular part of the agent’s inner structure, such as Reason or Appetite, from which it issued—in contrast to Watson’s early view. Other things being equal, the greater the morally relevant psychological factors underlying an action are “integrated into an agent’s overall personality,” the greater the agent’s praiseworthiness or blameworthiness for the action.

On the Whole Self theory, some motives, then, are privileged in representing the agent’s self, namely, the well-integrated ones. The privilege, however, is not in virtue of their issuing from a certain motivational source. Beliefs and desires are “well-integrated” to the extent that, first, they are “deep,” which is understood as a matter of an attitude’s being “a powerful force in determining the actor’s behavior”; in particular, beliefs are deep when they resist revision, and deep desires are those that tend to be satisfied with preference over “shallower” ones in cases of forced choice. Second, attitudes are well integrated to the extent that they do not oppose other deep beliefs or desires. An action, then, is well integrated just in case it results from well-integrated (deep and unopposed) beliefs and desires. Poor integration below some threshold, Arpaly and Schroeder maintain, is an excusing condition on moral responsibility.

The Whole Self view and the coherence account I have defended diverge in several respects. First, on the former, desires and beliefs generally figure into the right description of the self, not only beliefs formed with the aim of assenting to what is true (that is, acceptances) and desires critically processed with an eye...
toward what is good from one’s own perspective (that is, preferences). In other words, desires need not be formed by a process of reflective evaluation in order to count as internal to the true self, nor are beliefs adopted, for instance, only for the sake of argument, ruled out. Second, the Whole Self theory proposes a thinly described minimal requirement of non-opposition, rather than a fuller coherence condition of support with respect to worth against competitors. Third, Arpaly and Schroeder deem all beliefs resisting revision “deep,” not only those coherent with one’s critically evaluated reasons. Fourth, on the Whole Self view, desires are deep to the self when they are often satisfied and win out in cases of conflict with other desires.

On the Whole Self account, then, the core self is defined by what we often do and what we often assent to without conflict with other often-satisfied or persistently held beliefs. Consider the last point of disagreement first. Although the Whole Self theory takes desires to be deep when they tend to be satisfied over other desires—that is, when they are “a powerful force in determining the actor’s behavior”—notice that often-satisfied desires may in fact be quite shallow to the self, both from a third-person perspective and from the point of view of the agent. Recall the desires for repeated self-injections on the part of Frankfurt’s unwilling addict. By hypothesis, the unwilling addict hates his addiction, struggles against it, and wishes to overcome it, preferring never again to act on the desire to inject the drug. This latter preference is not satisfied, and hence on the Whole Self view, it is ruled out from counting as an aspect of who the addict really is. But surely this is counterintuitive to our understanding of an unwilling addict: Although such an addict regularly acts on the desire for the drug and may not regularly, or ever, act on conflicting desires (to refrain, to get help, to work out, instead), he is not “in his heart of hearts” a lover of drug taking. He is, rather, a slave to his addiction. The unwilling addict experiences grief, perhaps even revolt and despair, over his drug-taking, and given this fact, he would likely consider it inhumane for us to insist that the desire for drug-taking is a deep aspect of his self, definitive of his true psychological identity. (“But I hate that desire,” he might protest, “I take the drug in spite of myself.”)

I have suggested that when we act from our coherent preferences, we do not oppose or hinder ourselves, whereas, by contrast, when we act “in spite of ourselves,” we act on unendorsed desires and beliefs, not ones we are prepared to defend with reasons against competitors. Drug-taking on the part of the unwilling addict seems to frustrate, and not to express, his true psychological identity. It is difficult to see how the Whole Self theory could account for this fact, and so it is puzzling how it could explain the differences in autonomy between an unwilling and a willing addict. Even though regularly acted upon and not in conflict with other regularly satisfied desires, desires may be trivial and superficial to the self.

Furthermore, not all beliefs that resist revision should count as deep aspects of the agent’s psychology. This is the case because a belief may resist revision for any number of reasons, only some of which confer authenticity on the belief. For instance, a belief may be persistently retained due only to the influence of hypnosis, or to stubbornness, or to laziness in changing habit, or to indoctrination. Not all resistant beliefs are retained because the agent, having subjected them to sus-
tained critical scrutiny, now has good reasons for maintaining them. A belief one retains out of fashion, post-hypnotic suggestion, or laziness seems not deep but shallow to the true self.

Consider one final alternative: on an entirely different approach to the nature of the self, psychological elements count as authentic just in case, when violated, transgressed or unsatisfied, psychological breakdown ensues. The self’s central elements, on this sort of approach, are understood in terms of tendencies or dispositions to psychic disintegration. This type of view is problematic. Some persons are naturally more prone to mental or psychological breakdown than are others, without this showing that they have larger or more expansive true selves. For any variety of reasons including biochemical ones, depression easily affects some individuals, for instance; psychosis and other sorts of personality disintegration affect others, while still others are psychically more stable, relatively unfazed by disappointment and loss. Those less prone to breakdown do not necessarily have smaller selves than those more prone to it. And those who are so emotionally stable that they do not—and will never in any circumstances—psychologically break down, do not have non-existent selves. A particular value or belief or preference might be central to one’s moral or psychological self—for instance, one’s preference that one’s children live full and healthy lives—yet one may have such positive psychological and other resources that, were this desire unfulfilled, one would not disintegrate. The test for how much one cares about something—its centrality to one’s self—should not be one’s tendency toward psychic death.

7. CONCLUSION

A basic insight that Frankfurt has sought to elaborate and refine through the preceding three decades is that acting autonomously is acting on a motivation with which one identifies. Some psychological states, even when they motivate one to act, are experienced as foreign or alien. Though they are aspects of one’s psychological history, since one is alienated from them, they are not, in some important sense, one’s own. I have here proposed particular understandings of the notions of alienation and identification, working to incorporate what is valuable about the original Frankfurtian approach into a different theory of the self. I have explored the question of whether or not coherence in any sense among certain of one’s psychological states is required for autonomous action and have refined and illustrated the notion of coherence at issue. The result is a coherence theory of personal autonomy incorporating procedural elements.40

40. I gratefully acknowledge support for this work provided by a faculty research assignment from the College of William and Mary and by a Robert F. and Sara M. Boyd research fellowship. I have benefited from discussions with Nomy Arpaly, Randolph Clarke, Keith Lehrer, and Alfred Mele.