PHILOSOPHICAL TOPICS
VOL. 41, NO. 1, SPRING 2013

Values, Agency, and Welfare

Jason R. Raibley
California State University, Long Beach

ABSTRACT. The values-based approach to welfare holds that it is good for one to realize goals, activities, and relationships with which one strongly (and stably) identifies. This approach preserves the subjectivity of welfare while affirming that a life well lived must be active, engaged, and subjectively meaningful. As opposed to more objective theories, it is unified, naturalistic, and ontologically parsimonious. However, it faces objections concerning the possibility of self-sacrifice, disinterested and paradoxical values, and values that are out of sync with physical and emotional needs. This paper revises the values-based approach, emphasizing the important—but limited—role consciously held values play in human agency. The additional components of human agency in turn explain why it is important for one's values to cohere with one's fixed drives, hard-wired emotional responses, and nonvolitionally guided cognitive processes. This affords promising responses to the objections above.

I. INTRODUCTION

Theories of personal welfare aim to explain the nature of personal welfare or well-being. They tell us what it is for a life to go well (or badly) for the one who lives it or, equivalently, what it is in virtue of which a human life scores high (or

low) in welfare-value. Welfare is intimately connected with the ordinary concepts of benefit and harm, as well as prudence, self-interest, and beneficence. A person's welfare-level is the thing that is increased when a person receives a direct or noninstrumental benefit; the welfare-value of a person's life increases when a person receives an overall benefit. Welfare is also the thing that is diminished when a person suffers a direct or noninstrumental harm; the welfare-value of a person's life decreases when they receive an overall harm. If something is one's interest, it is thereby at least likely to augment one's welfare. Similarly, prudential or self-interested reasons for action would be reasons to promote or advance one's own welfare (or to do what can be rationally expected to promote or advance one's welfare). Reasons of beneficence, by contrast, are reasons to promote the welfare of others (or to do what can be rationally expected to promote the welfare of others). Utilitarians typically believe that aggregate or group welfare is the thing that morally ought to be maximized. Rational egoists hold that it is one's own welfare that one rationally or all-things-considered ought to maximize. Welfare-value is also used to explain which traits are virtues in some versions of virtue ethics; these theories define virtues as traits that one needs to live a life high in personal well-being.

Welfare-value must be carefully distinguished from other evaluative and semi-evaluative properties that can be instantiated by human lives. Such properties include happiness in the contemporary psychological sense, moral virtue, meaningfulness, various forms of aesthetic value, and enviableness. None of these properties is simply *the same property* as welfare-value. Although the term 'happiness' is sometimes used as a synonym for welfare or well-being, it is used by most contemporary speakers of English to pick out either (a) a temporary mental state (as in "Susan is eating her favorite food; look how happy she is!") or (b) a complex, diachronic emotional condition (as in "After years of struggle, Jessica is finally happy in both her work and home life"). These are both distinct from welfare or well-being, though a person could not be exceptionally well-off if he or she were mostly unhappy in either of these senses (Raibley 2012; cf. Haybron 2008). It might be that moral virtue always improves lives that contain it, but the welfare-value of a life cannot simply be equated with its degree of moral virtue: otherwise, it would be impossible for the morally wicked to flourish, and it would be impossible for the morally virtuous to lead lives that went badly for them. Similar things are true of meaningfulness in both its objective and subjective forms. Meaningfulness may be intimately related to personal well-being, but the two things are not the same thing, or else a richly meaningful life would be entirely beyond improvement when it came to personal welfare. Furthermore, it is possible for lives to exhibit narrative unity—or a kind of dark beauty, or other aesthetic features—without going well for those who live them. Finally, a broad range of features—including both forms of happiness just mentioned, moral virtue, subjective or objective meaningfulness, aesthetic beauty, and even the time and place at which it is lived—could serve to make a life more enviable, without serving to raise the level of welfare of the one

who lived it. This is because people care about things other than welfare, and so they may envy others if their lives contain these things. None of this, of course, is to deny that these things—happiness, virtue, meaningfulness, etc.—might have a bearing on individual welfare; it is merely to say that these are all distinct properties of human persons or lives.

Many of the leading theories of welfare specify basic bearers of noninstrumental welfare-value: they specify states or episodes that improve individual human lives, simply in-and-of-themselves, or irrespective of their consequences or results. The greater the number of the relevant states and episodes in an individual's life, the better it is for them. Popular theories of welfare that have this atomistic structure include hedonism (e.g., Feldman 2004; Crisp 2005), desire-satisfactionism (e.g., Brandt 1979; Griffin 1986; Heathwood 2005), and the Objective List Theory (e.g., Parfit 1984; Rice 2013). Some leading theories do *not* have this sort of atomistic structure. For example, leading versions of life-satisfactionism (e.g., Sumner 1996) explain the welfare-value of a person's life at a time in terms of the attitudes that the person would take to their life as a whole at that time. On this theory, the welfare-value of a life might change dramatically from one instant to the next. The form of preferentism advanced by John Rawls in *A Theory of Justice* is another nonatomistic approach (Rawls 1971). Rawls's theory is holistic; it specifies the ideal life for an individual by considering which of various possible lives the person would rationally prefer. It then measures the person's welfare in actuality by comparing their actual life to this ideal life.

All these theories of welfare face serious problems, as do closely related psychological models of welfare, happiness, and 'subjective well-being' (e.g., Diener et al. 1985; Diener and Biswas-Diener 2008). While the literature on these problems is extensive—and many interesting emendations to these theories have been proposed—the following summary of leading difficulties will serve to illustrate the promise of a competing approach, the values-based theory of welfare.

Hedonism—which holds that positive welfare consists in the accumulation of episodes of pleasure—construes welfare purely as a mental state. On this theory, if a person took pleasure in the 'fact' that they had achieved various things—when they had actually achieved none of these things—then the welfare-value of their life would be equal to an otherwise similar person who actually *did* achieve those same things (Nozick 1974; Nagel 1979; Kagan 1994). Many have found this result difficult to accept.

In its most basic form, desire-satisfactionism states that one's welfare is increased when one desires that some situation obtain and this situation actually does obtain, while one still desires it; the benefit one receives is proportionate to the strength of the relevant desire. This theory does not suffer from the problem just described in connection with hedonism, but many versions of this view imply that the satisfaction of urges (say, to knock down icicles) and compulsions (say, to wash one's hands) is just as beneficial as the realization of one's most central cares and concerns (Kraut 1994; Raibley 2010). Some versions of this theory also imply

that desire-satisfaction is beneficial even when the subject does not know their desire has been satisfied, and even when the desire is for some outcome that seems quite remote and disconnected from one's experience. Suppose one talks briefly with a stranger on a train and learns that he is sick. If one forms a desire that he be cured, and if he actually is cured, then desire-satisfactionism implies that one's life is improved—even if one never learns that the stranger was cured (Parfit 1984; Griffin 1986).

The Objective List Theory holds that states such as knowledge, achievement, the appreciation of beauty, and true friendship benefit one, whether or not one desires or enjoys such states. This theory therefore appears to violate Peter Railton's widely accepted *internalism requirement*: "what is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive, at least if he were rational and aware. It would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him" (Railton 2003 [1986], 47). This problem has inspired some welfare axiologists to adopt hybridized versions of the Objective List Theory. On one such hybridized theory, it benefits one *to desire* and then *to get* knowledge, achievement, beauty, and friendship (etc.); on another, it benefits one *to enjoy* just these things. Like the simpler version of the Objective List Theory, these hybridized versions appear to lack unity. That is, they do not satisfactorily explain what these sources of positive welfare have in common with one another. Now, it could be answered that they all involve pursuit or enjoyment of the excellent or the intrinsically good (e.g., Adams 1998). But this response introduces its own difficulties: if the excellent or the intrinsically good is that which people have reason to pursue, promote, enjoy, and/or love for its own sake, irrespective of their own cares and concerns, then the response presupposes the existence of a 'queer' form of objective value (Mackie 1977).

Life-satisfactionism states that one's life is going well for one to the degree that (a) one judges that it is satisfactory and (b) one is emotionally satisfied with it. But what if a person is simply very difficult to satisfy? Does this mean that their life is going badly for them, no matter how much pleasure, knowledge, achievement, friendship, and beauty it contains? A related problem is that even the most refined versions of life-satisfactionism seem to imply that it is not possible for a person to be mistakenly dissatisfied with their life, provided that their judgment about it is informed, autonomous, and based on stable and affectively appropriate goals and attachments (Raibley 2010). Another problem is that the theory seems to imply that one lacks a welfare-level altogether if one never pauses to take stock of one's life, and so never renders any judgment about it, one way or another (Feldman 2010). Additionally, judgments of life-satisfaction seem overly susceptible to framing effects, social comparisons, and assorted cognitive biases (Schwarz and Stark 1999; Haybron 2008).

Finally, the form of whole-life preferentism advocated by Rawls also faces several important problems. On this theory, the ideal life for a person is the life that they would prefer on the whole if they were fully informed and possessed of full

deliberative rationality. This proposal is difficult to evaluate. It is not entirely clear what either full information or full deliberative rationality come to, or whether they would 'improve' a subject's preferences in the intended way (cf. Velleman 1989; this problem also affects the rational or ideal desire-satisfaction proposed in Brandt 1979). Additionally, at what time in a person's life are we to apply this procedure? At age 18, or age 50, or age 85? It seems that we might get different results in each case, because preferences might be impacted by factors other than information and deliberation, including natural changes in a person's sensibilities or appetites brought on by age. Most importantly, even if this approach correctly identifies the ideal life for an agent, why should we think that the second-best life for the agent would—in all its specific and concrete features—resemble this ideal life to a high degree? It seems that there might be several very good possible lives for a person that are, in concrete terms, all markedly different from the ideal life for the person (Raibley 2012b).

The problems with all these theories have motivated interest in the *values-based theory of well-being.* This theory is implicit in the works of several writers (Frankfurt 1988; Gaus 1990; Anderson 1993; Copp 1995; Hubin 2003; Tiberius 2008; cf. also Griffin 1986; Goldman 2009), but it has received explicit formulation only more recently (Raibley 2010; Tiberius and Plakias 2010; Tiberius and Hall 2010; Haybron and Tiberius 2012). Its guiding idea is that it is directly beneficial for people to have and knowingly to realize their *values*, including goals, activities, relationships, aims, ideals, and principles that they care about for their own sake.

For the advocates of values-based theories of welfare, 'valuing' denotes a psychological attitude distinct from desire. Valuing is not a purely cognitive attitude; even if valuing a thing typically involves also believing that it is good in some sense, valuing is not merely a doxastic state. Nor is it necessary for a person to believe that something is objectively good, in order for them to value it. Rather, valuing is in significant part affective or emotional. It is always partially constituted by some diachronically stable and noninstrumental pro-attitude, such as desiring, enjoying, liking, loving, caring, or esteeming. But a stable and noninstrumental pro-attitude is not yet constitutive of valuing, unless the agent also *stably identifies* with the pro-attitude in question. To stably identify with the pro-attitude, the agent must be disposed to take it to be representative of who they are and who they want to be. In addition, the pro-attitude must inform or structure their emotional responses and practical deliberations. Importantly, the pro-attitude must be "normative from [the agent's] point of view," which means that the agent must be disposed to take the pro-attitude to be justificatory of actions (cf. Bratman 1996; Haybron and Tiberius 2012). A person's values are not the things that are *notionally* important to them, or the things that they believe to be objectively worthwhile. Rather, their values are the objects of egosyntonic pro-attitudes that serve to guide their actual behavior, e.g., by figuring in their practical deliberations. Finally, values must be adequately informed if they are to be correctly attributed to agents: a person does not really value a given item if they are ignorant of its true nature.

Values, on this approach, are realized in different ways: activities are pursued and enjoyed, relationships are nurtured or maintained, goals and aims are achieved, ideals are promoted and preserved, principles are observed, objects are cherished or preserved, and persons are nurtured or protected. But the life high in personal welfare involves engaged activity of some such form on behalf of (or guided by) one's values.

While the values-based approach to welfare can be developed in different ways, one promising formulation takes its cue from versions of aim-achievementism, a closely related theory (Scanlon 1998; Keller 2004; Keller 2009). Developed in this way, the theory would say that it is directly good for a person when they knowingly realize one of their current values through their own activity (call this an episode of 'value-realization'). The welfare-value associated with such an episode of value-realization is proportional to the importance of the value to the subject, during its realization. Next, it is directly bad for a person to fail in the realization of their current values (this would be an episode of 'value-frustration' or 'value-impairment'). Such an episode's negative welfare-value is also proportional to the importance of the value to the subject. Finally, the degree to which a segment of a human life (up to and including a whole life) is good for a person depends mainly on the welfare-values of the episodes of value-realization and value-frustration that it contains ("mainly," because many advocates of the approach allow that hedonic and emotional ills, such as nociceptive and attitudinal pain, anxiety, and compression, are also directly bad for a person).

The values-based theory of well-being corrects some of the problems noted in connection with the theories mentioned earlier. It does not construe welfare as a mental state: welfare is an objective condition that requires bringing about a sort of match between one's pro-attitudes and the world. Nor does it treat *every* impulse, urge, or whim as important to welfare; it gives pride of place to values. It implies that one's welfare is increased only when one plays an active role in realizing one's values: it neither benefits nor harms one when a desire that does not prompt action—e.g., the desire that the stranger's illness be cured—is satisfied, even if one knows it has been satisfied. The theory is also compatible with the idea that people can be dissatisfied with their lives when they rationally ought not to be. Similarly, the theory assigns welfare-levels even to those who have never paused to assess their own lives. Unlike the Objective List Theory, the values-based theory satisfies Railton's internalism requirement, is highly unified, and avoids invoking any metaphysically 'queer' sorts of value. Finally, unlike Rawls's theory, it does not invoke the problematic concepts of full information or deliberative rationality, nor does it measure all lives by their distance in concrete detail from the ideal life for a given agent.

The theory may have additional advantages. First, many people attach great importance to well-being as a life goal. Furthermore, they believe that achieving a high level of personal welfare is no trivial task. Hedonism, desire-satisfactionism, and life-satisfactionism are in some tension with these ideas. Pleasure is not so hard to obtain. Desire-satisfaction can be ratcheted-up by cultivating

desires for things one already has or things that are very easily obtained. Life-satisfaction can be achieved by lowering one's standards. But since it is more difficult to manipulate one's own values—to alter what one really cares about—it will also be difficult to find a shortcut to personal welfare if the values-based theory is true. Second, this same theory allows for indirect but important connections with reasons for action and meaning in life. Many believe that self-interest or prudence normally generates reasons for action. Since it does not seem that prospective pleasure and visceral urges generate such reasons (Copp 1995), theories of welfare such as hedonism and desire-satisfactionism are in tension with the thought that one normally has *some* reason to do what would most advance one's welfare. However, Humeanism, which is a leading approach to reasons for action, can be formulated in terms of values (Hubin 1999; Tiberius 2000; cf. Goldman 2009). If a values-based form of Humeanism is true, and if the values-based theory of welfare is also true, then prudence will normally generate real, normative reasons for action. Furthermore, because of the way values are analyzed on this approach, it is plausible that a person will necessarily find some (subjective) meaning in the successful realization of their own values. And so on the values-based theory, a life high in welfare will necessarily also be a (subjectively) meaningful life. This accords with the idea that a life couldn't go well for the one who lived it if they experienced it as meaningless. By contrast, it seems difficult to forge a connection between meaningfulness and thinner attitudes, such as enjoyment and desire: a life could be very pleasant indeed—or involve many gratified urges—without being experienced as meaningful.

Despite all the advantages mentioned here, the values-based theory faces several important objections. The remainder of this paper focuses on two sets of objections, both of which relate to the concept of valuings.

The first set of objections is familiar from discussions of desire-based theories of welfare. Values—or more precisely, valuings—share two important features with desires: they are conative or motivational, and they involve a 'world-to-mind' direction of fit. Consequently, like desire-based theories of welfare, values-based theories are thought to have a variety of paradoxical and counterintuitive consequences. First, some say that these theories cannot adequately distinguish between selfish and disinterested values. Second, such theories appear to render self-sacrifice conceptually impossible. Third, these same theories make certain values (e.g., one's own ill-being) paradoxical that do not seem paradoxical.

The second set of objections relates to the nature of values, in particular the stable identification requirement discussed above. If stable identification requires reflective endorsement of one's pro-attitudes—or at any rate, having 'higher-order' attitudes toward one's 'first-order' pro-attitudes—there is the possibility that the values-based theory will not pay adequate attention to ordinary desires or to various nonconscious or subpersonal processes and states that also have a direct bearing on personal well-being. Consequently, while the move from 'first-order' pro-attitudes to values may solve some problems, it may introduce others.

It will be shown that the first set of objections can be partially addressed without making any drastic changes to the values-based theory. However, the amendments to the theory that are required to address the second set of objections also help us more fully to address the first. It is ultimately argued that a holistic, agency-based theory of welfare can preserve the advantages of the values-based theory while answering both sets of objections.

II. DISINTERESTED VALUES, SELF-SACRIFICE, AND PARADOXICAL VALUES

As just noted, valuings are conative or motivational. They guide human action—perhaps they can even cause it. Relatedly, and like other pro-attitudes, they involve a 'world-to-mind' direction of fit—when you value something, you are thereby disposed to change (or in some cases maintain) things so that the world matches your ideal. These two features of the theory seem to have several paradoxical and counterintuitive implications, viz., that there is no distinction between self-serving and disinterested desires or values, that self-sacrifice is impossible, and that it is impossible to have a deep, reflective, egosyntonic desire for one's own ill-being.

2.1 THE OBJECTION FROM DISINTERESTED DESIRES

The objection from disinterested desires is closely related to the case of the stranger's cure, discussed above. As applied to the values-based theory, it runs as follows. If the values-based theory is true, then all values are in a sense self-serving: any value is necessarily such that its achievement directly benefits the agent who achieves it. This conclusion is somewhat surprising. Suppose a man who lives in the United States desires (and stably identifies with his desire) that the clear-cutting of the Brazilian rainforest be stopped. Suppose it is stopped, while he still exists and still cares about this outcome. Suppose also that he *knows* that it has been stopped, but that he does not live to enjoy any indirect benefits from its being stopped in the form of, e.g., improved air quality. If the values-based theory is true, some might say, then this event benefits the man, and so the value in question is a self-interested value. But, it is objected, what happens in the Brazilian rainforest can have no direct impact on this man's level of well-being. And, what is more important in the present context, caring about the preservation of the rainforest is a paradigmatic example of a selfless or disinterested value. Therefore, the values-based theory is false.

This instance of the objection would be based on a misunderstanding of the values-based theory. Earlier, it was stated that, in order for a pro-attitude fully to count as a value, the pro-attitude must inform or structure one's emotional responses *and practical deliberations*. It was specified that values are neither the

things that are merely notionally important to a given agent, nor the things that the agent would agree are objectively worthwhile. For it is possible to see some outcome or situation as valuable, good, worthwhile, or morally correct without incorporating that outcome or situation into one's system of personal ends. The values relevant to the values-based theory are values that guide the agent's actual deliberation and action. As the case was described, the man desires (and identifies with his desire) that the clear-cutting of the rainforest be stopped. But he never actually does anything about it—perhaps he even believes that there is nothing he *can* do about it. Protection of the rainforest is not something that he values in the relevant sense. Furthermore, the man plays no part in bringing it about that the clear-cutting was stopped. If he does not causally contribute to the attainment or protection of a value through his own activity, then he does not *realize* the value, either. (Note that the theory does not require that one be *solely* responsible for the value's realization, but it does require that one make *some* causal contribution.)

Still, it might be thought that the problem has not entirely gone away. Suppose protecting the rainforest *is* a value by reference to which the man structures his own activities: suppose he desires that he contribute to the salvation of the rainforest. Suppose he stably identifies with this desire, and that he forms intentions based on this desire. Furthermore, suppose it *is* partly due to his own efforts that the clear-cutting is stopped (perhaps he raises money for charities that make some causal contribution to this outcome). Suppose he also knows that the outcome is attained. It might still be objected that (a) the outcome does not directly benefit the agent, i.e., that the achievement of this valued goal by the agent does not directly improve the welfare-value of his life, and (b) that even if it *did* benefit the agent, the value in question is not a self-interested or selfish value.

The values-based theory does inevitably imply in this case that the actualization of this outcome benefits the agent to some degree. However, the theory states that its direct welfare-value for him will depend on how much he cares about the outcome, and that its overall value for him will depend on the degree to which his other values are realized or not realized as a consequence of his pursuing this end. The theory could also be modified in minor ways so that the value of this action also depends on how much of his activity he devotes to it, or his precise causal contribution to the outcome (cf. Portmore 2007). For these reasons, it is not clear that the values-based theory's implications, here, are really that counterintuitive. Consequently, objection (a) is not very strong.

Next, with respect to objection (b), even if the outcome directly benefits the agent to some degree, this does not require us to say that the relevant value was selfish or self-interested in the pejorative sense. There are a variety of ways of characterizing selfish or self-interested values without saying that they are always irrelevant to one's welfare when achieved. Perhaps the main thing that determines whether a value is selfish or not is whether the person holds and pursues it *because* or *in virtue of the fact that* they care about their own welfare (or pleasure), as

opposed to holding and pursuing it for some other reason (e.g., because they care about the welfare or pleasure or improvement of others, or about beauty, or about some other moral or aesthetic ideal). Alternatively, whether a value is selfish or self-interested in the pejorative sense may have to do with whether its realization will bring pleasure and satisfaction primarily to the agent, or whether it will also bring significant pleasure or satisfaction to others. The point is, there are other plausible ways of understanding the distinction between selfish and disinterested values that do not require the assumption that the realization of disinterested values cannot benefit one.

Now, it might appear as though a false theory of welfare, such as simple hedonism, is 'built-in' to the ordinary distinction between selfish and disinterested values. If this is the case, this certainly does not prove that hedonism is true (though it may indicate that hedonism is widely assumed to be true). Nor would this entail that we ought to get rid of (or revise) our distinction between selfish and disinterested values, should the values-based theory prove correct. For even if the values-based theory is correct and hedonism is false, there may be important moral or practical reasons to distinguish people who are motivated primarily by their own pleasure or satisfaction from people who care intrinsically about things beside their own pleasure and satisfaction. Accordingly, it may make sense to distinguish between values born of these two motivational tendencies.

2.2 THE OBJECTION FROM SELF-SACRIFICE

The objection from self-sacrifice was famously stated by Mark C. Overvold (1980): "[I]f we identify an agent's self-interest with what he most wants to do, all things considered, it becomes logically impossible that there ever by a genuine instance of self-sacrifice" (117). This objection targets certain desire-satisfactionist theories of welfare. These theories apparently involve the idea that, if some outcome is most preferred by an agent at a time—or if there is at least no other outcome that the agent prefers to it—and if this agent then performs an action that brings about this outcome, then this action is *maximally self-interested*, i.e., it advances the agent's welfare at least as much as any other action that the agent might have performed. But, it is thought, for an act of self-sacrifice to occur, the act must not be maximally self-interested. And so any act that produces an outcome most preferred by an agent at a time cannot be an act of self-sacrifice.

It might be thought that this same problem arises for the values-based theory in something like the following way. Suppose that a 29-year-old woman belongs to the privileged class in her own country. While she is lucky to enjoy some degree of wealth and comfort, she is a political dissident, and she values above all other things the democratic reform of the political system in her country. In her judgment, an act of public protest at a critical moment will best realize this value. However, this act of protest will predictably result in her spending years in prison or under house-arrest. She realizes this, and she performs the act, anyway.

Intuitively, her act is an act of self-sacrifice. But the values-based theory, it might be argued, cannot classify it as such: this theory must classify the act as maximally self-interested, since it was expressive of the woman's deepest values. But (it is thought) a maximally self-interested act cannot simultaneously be self-sacrificial!

Things are not quite so simple. As Chris Heathwood points out in his recent paper, "Preferentism and Self-Sacrifice," Overvold's objection assumes that we ought to "determine how good an outcome would be for a person by looking to the person's desires about the outcome," when instead, a better conative approach tells us to determine "how good an outcome would be for a person by looking at how well-satisfied the desires *within the outcome* would be. The best outcome for the person is the one that best satisfies the desires she will have if it comes about" (Heathwood 2011, 20; emphasis mine). The values-based theory stated above *does* imply that, if the dissident values these reforms at the time she achieves them, this is directly beneficial to her to some degree. However, the theory does *not* imply that this act was maximally self-interested. The maximally self-interested act would be *the one that conduces to the greatest net amount of value-realization over the woman's life.* But the case has not been described in a way that guarantees the act has *this* feature.

The point may be illustrated more vividly if we slightly modify the original example. Suppose that our political dissident values democratic reform above all other things. Suppose that her act of public protest will best accomplish this reform, but that the predictable consequence of this protest will be her own death. Suppose that she knows this and performs the act, anyway. She is then executed by the state. The theory of welfare Overvold apparently has in mind would, implausibly, count this act of protest as maximally self-interested. But this only shows that Overvold is targeting an inferior version of desire-satisfactionism. A better desire-based approach would say that the maximally self-interested act for the dissident is *the one that involves the greatest sum-total of desire-satisfaction over the dissident's lifetime*. No matter how much she desires reform, it seems likely that her net amount of desire-satisfaction would be higher if she were to do something else—something compatible with her living a normal human lifespan. From the fact that an agent values some outcome above all others at a time and chooses to act so as to actualize this outcome, it does not follow that this outcome is best for the agent, or that her act is maximally self-interested. These things will depend on what values the agent will come to have later on, if she performs this act, now, and on the degree to which these are eventually realized. Therefore, the values-based theory allows for self-sacrifice in this sense: an agent might choose what is worse for her on the whole.

Still, going back to the case where the penalty is imprisonment and not execution, the question remains: *What if* the dissident genuinely does value political reform to such a high degree—and in such a wholehearted way—that she feels she simply *must* do her part to achieve reform, even if it means her imprisonment? *What if* she cares much, much less about her happiness, her physical and psychological health, her relationships with her friends and family members, and

all the other things people usually care about? Unrealistic though it may be, *what if* she will simply cease to care much about anything at all if she fails to protest at this critical moment? In short: *what if* it actually would maximize lifetime value-realization for her to go through with the act of protest?

The first thing that should be said in reply is that we need not classify the act as self-sacrificial in order to account for its moral merit. Neither the degree to which it promotes moral goodness—nor the degree to which it satisfies an imperfect duty to see others' needs and projects as one's own—is affected by whether the act is counted as self-sacrificial or not. It is not even clear that the act must be self-sacrificial in order for it to express the virtues of beneficence and justice. Once this is noted, the defender of the values-based approach might say that it is not clear, in this situation, that the dissident's act really ought to be counted as self-sacrificial. If she really cares this much about political reform, it seems that the act is expressive of her practical identity—it is more 'true' to her own self than any other act she might perform.

Once again, though, there is another assumption that can be brought into question, namely, the assumption that acts of self-sacrifice cannot be welfare-optimific. A person can be directly harmed even when the act that harms the person promotes their welfare more than any available alternative. For example, if the welfare-optimific action maims or disfigures the agent, it is still intuitive to speak of it as a 'harmful' action, despite the fact that it produced the best possible outcome in the context (Shiffrin 1999; Harman 2009). Similarly, it seems possible for self-sacrifice to occur even when the self-sacrificial act is welfare-optimific, provided that the act simultaneously involves or directly results in the loss of important basic welfare-goods, such as happiness, physical and psychological health, and sociability. In section 5 below, we will develop a theory according to which there are indeed such basic welfare-goods; doing this will further fill out and justify this approach to the problem of self-sacrifice. But the basic idea is simply that self-sacrificial acts *can* be welfare-optimific, provided that these acts involve major losses like those that the dissident will suffer in prison.

2.3 THE OBJECTION FROM PARADOXICAL VALUES

Perhaps a deeper worry concerns those who are consciously and intentionally self-destructive. To take the most extreme case, consider someone whose only desire is to fare badly, overall.[1] The person identifies with this desire, forms intentions based on this desire, and takes acts to be justified to the degree that they tend to satisfy this desire. In short, this person values their own overall ill-being. If the values-based theory is true, then the realization of this value is paradoxical. For if the agent successfully realizes this value, then the agent is *both* faring well (because realizing this value) and not faring well (because the value in question is ill-being, and in order for it to be realized, the agent must be faring badly). Ben Bradley (2009) describes the problem for the desire-based theory of welfare as follows,

addressing the suggestion that such paradoxical values are no more worrisome than the liar paradox:

> The desire to have one's life go badly [on the whole] is not like the liar sentence, nor is it like the desire to have one's desires frustrated. *It is not transparently paradoxical.* It seems like an unproblematic desire. It is paradoxical, and has liar-like features, only given a particular theory of welfare … there is at least one desire [i.e., the desire for one's own overall ill-being] that does not seem paradoxical, but in fact is paradoxical if, but only if, [the desire-based theory] is true. (2009, 40; emphasis in original)

In response to this objection, the values-based theorist should simply embrace and defend the idea that it is paradoxical to value one's own ill-being: just as it is impossible to satisfy one's desire that all one's desires be frustrated—just as it is impossible to believe that all of one's beliefs are false—it is impossible exclusively to value one's own ill-being and then realize this value. As Bradley notes, valuing one's own ill-being is not *transparently* paradoxical. But if the nature of welfare is not itself transparent—so that its deep nature need *not* be apparent to those who are perfectly competent users of the welfare-vocabulary—then paradoxes involving welfare need not be transparent, either. In particular, if welfare is something like a natural kind, then we should not expect its nature to be transparent. So, for example, if individual well-being is a "homeostatic property cluster," as Richard Boyd suggests, then its nature will only be discoverable by way of investigations that are at least partially *a posteriori* (Boyd 1988 fn. 2; cf. Kornblith 2002, which pursues the hypothesis that knowledge is also akin to a natural kind). However, after one has pursued the relevant investigations and been presented with the relevant evidence—including the problems with traditional theories—the paradoxical nature of realizing the goal of one's own ill-being becomes apparent.

However, part of the intuitive appeal of the objection is that it *does* seem possible to pursue and achieve one's own ill-being—it seems that people actually succeed in doing this. The above reply to the objection would therefore be more satisfactory if we could also provide a model showing how a person might value their life going overall badly and achieve this (even though this particular episode of value-realization contributes positively to the value of their life). Such a model is provided, in effect, by those who insist that, even if a person desires that their life go badly, they cannot—if they are a human agent—help having many other desires besides this one. These philosophers go on to argue that a person's ill-being can be achieved by the frustration (over time) of a sufficient number of these many other desires. Those who favor pluralistic, holistic, and multifactor models of welfare that treat value-realization as *one*—but only *one*—direct determinant of well-being can provide similar explanations of the possibility of self-sacrifice. In particular, on the holistic modification of the values-based approach presented below, persons can satisfy their deep desire for ill-being—even if realizing this value contributes to their welfare—provided that they act to undermine their health, happiness, longevity, or other basic welfare-goods in sufficiently radical

ways. This modified theory also enables us to say precisely what it is that the person who values ill-being values, and how this value might be realized without them faring well, even if realizing this value adds directly to their welfare.

III. HIGHER-ORDER PROBLEMS

In addition to the better-known objections to the values-based approach just discussed, this theory faces additional challenges from recent work in moral psychology. As already noted, the values-based theory attempts to forge more credible constitutive connections between well-being and pro-attitudes by focusing on valuings, which are pro-attitudes such as wanting, liking, and loving that are accompanied and ratified by attitudes and dispositions concerning those very attitudes. It was stated that values paradigmatically involve lower-order pro-attitudes that are (a) adequately informed and (b) stable. Also, the agent should (c) identify with the pro-attitude—which typically involves approving of the attitude itself and being un-conflicted about it, so that the agent does not disapprove of it at any higher "level." The agent should also be (d) disposed to treat the pro-attitude as justificatory of their actions. (Degenerate and borderline cases are still possible, but pro-attitudes count as values to the degree that they exhibit all these features.) The thought was that, by focusing on these attitudes, as opposed to urges, whims, or attitudinal pleasures, the values-based theory might capture welfare, itself, as opposed to contentment or episodic happiness or good mood. However, relying on pro-attitudes that have these additional features also introduces at least three new problems. These problems are described in the remainder of this section.

3.1 MERE DESIRES

One potential problem with the values-based approach is that it implies that the satisfaction of mere desires is of no benefit whatsoever. Consider this illustration. Suppose a person is walking through a public market and sees a vendor with a large crate of nice-looking apples. Suppose the person spontaneously desires to purchase and eat one of these apples, and does so. Would the gratification of this desire not be of at least some marginal benefit? The values-based theory may be well motivated, insofar as an adult human life could not go exceptionally well for the one who lived it if this life was entirely devoid of values and consisted merely in the gratification of desires. Still, the satisfaction of mere desires ought to count for something, and the values-based approach as stated earlier is therefore incomplete.[2]

3.2 INAUTHENTIC VALUES

Some philosophers have recruited higher-order pro-attitudes similar to the values invoked by the values-based theory of welfare to differentiate between voluntary

and nonvoluntary action, or between desires that have normative authority and those that lack it (cf. Frankfurt 1988; Korsgaard 1996). But other philosophers have strenuously criticized the deployment of higher-order pro-attitudes for such purposes (cf. Arpaly 2003; Levy 2007; Kornblith 2012). These critics have argued that identification with an attitude, even if the identification is informed and reflective, does not guarantee that the attitude will have the right sort of normative authority. They have also argued that emphasizing consciously held evaluations presupposes a problematic model of the mind—a model on which all the contents of the mind are transparent and on which a person's conscious or avowed estimations of things can be taken at face value. Some of the concerns raised by these critics seem relevant in the current context, too.

For example, Nomy Arpaly describes several individuals whose second-order attitudes are of questionable authority. Consider these two cases:

> Imagine a person, Lynn, who discovers that she is a lesbian and is deeply disturbed by that discovery. Her homosexual desires conflict with her values and her sense of her identity. She does not want her desires to motivate her into action under any circumstances—the very thought scares her more than anything else. ... If ... she were to read the moral psychology literature and believe its claims, she would probably conclude that she was right and her homosexual desires are not truly her own. (Arpaly 2003, 16)

Arpaly also quotes the following example from Christine Korsgaard:

> "I see a piece of cake in the fridge and feel a desire to eat it. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse does not dominate me and now I have a problem. Is this desire really a reason to act? I consider the action on its merits and then decide that eating the cake is not worth the fat and calories. I walk away from the fridge, feeling a sense of dignity" (Korsgaard, 1996, p. 100).

Arpaly then comments:

> This could be ... the inner monologue of an individual with severe anorexia nervosa, weighing under eighty-five pounds ... a woman who ... appears to her friends (or even to her future self, after having recovered from her anorexia or her irrational dieting) to be a person who *is* in fact at the mercy of her desires, or at the mercy of what are commonly called her "emotional issues." ... The anorectic is a potential challenge to contemporary moral psychology because she is a person who experiences her psyche in terms of self-control, as if there were something that was her, choosing between her desires on the basis of their merits, giving her control over herself, while we have good reasons to believe that unconscious desire or emotion moves her in a manner not characteristic of well-exercised practical reason. To the extent that moral psychology emphasizes the first-person perspective, it risks misunderstanding such cases. (Arpaly 2003, 17–18)

There are many ways of understanding the details in these cases, but one possible way of understanding them does seem to threaten the values-based theory of welfare. Suppose Lynn desires to avoid same-sex sexual contact, that this desire is stable and adequately informed, that she approves of this desire and does not disapprove of it at any higher level, and that she takes this desire to be justificatory of action. Suppose, further, that she actively disapproves of all her homosexual desires, considers these desires to be foreign or alien, and does not take them to be justificatory of action. In this case, the values-based theory of welfare would apparently imply that it would benefit Lynn to avoid same-sex sexual contact and that it would be bad for her to embrace what we tend to think of as her true sexual identity.

Similarly, if the anorectic desires to maintain her unhealthy weight, and if this desire is stable and adequately informed, and if she takes it to be representative of who she really is (and to be justificatory of her actions), then the values-based theory implies that there would be significant direct benefit involved in the realization of this value, even if it comes at the price of reduced net value-realization over the anorectic's lifetime. In other words, it implies that it is *directly* good for her to some degree to restrict her food intake on a given occasion, though it also will probably count this same action as *overall* bad for her on the grounds that it frustrates the realization of other values. This result is very worrisome. For surely it does not directly benefit the anorectic to realize *this* value even at the time—surely it would be *intrinsically* (and not just *overall*) better for the anorectic to rid herself of the relevant desire than to satisfy it.

In both these cases, the agents' egosyntonic attitudes are not reliable guides to what would benefit them at a time, or to what it would be prudentially rational for them to do, or to what others who care about them thereby have reason to do for them. When we filter these agents' pro-attitudes so that only the ones with which they identify are admitted, we end up with precisely those attitudes that it would be *harmful* for these agents to act on in the here-and-now. Now, a defender of the version of the values-based theory stated above might object that these agents' desires could not *really* be adequately informed. If they really knew all the relevant facts, it might be said, they would surely give up these desires. But it seems dangerous to count on this, especially given that we do not want to invoke such a stringent information requirement as to reintroduce the problem of alienation (Velleman 1988). For this reason, other modifications to the theory seem necessary.

3.3 JOYLESS VALUES

In a related vein, Daniel Haybron has described several interesting cases where the achievement of self-professed values is accompanied by anxiety, nervousness, frustration, and other negative, burdensome emotions. Here is one:

> Consider also the case of Claudia, an attorney. Claudia ... prefers wealth and social status to happiness, and she found lawyering to be

> the most efficient means to these ends. She has succeeded, amassing great hordes of money, acquiring the finest luxuries, and earning the envy of her peers. But work at the kind of prestigious firm that meets her needs is, for her, stressful and emotionally unfulfilling. As a result, she is short-tempered, stressed out, anxious, and mildly depressed. She could be happy in other pursuits, such as teaching or painting, which she would see as perfectly meaningful and worthwhile. Yet such happy-making activities do not bring her the riches and social prominence she desires. She does not regret her choice, and would not accept a life of average means and standing for any amount of happiness ... her choice does not depend on errors in reasoning, factual ignorance, or thoughtlessness: these are her values. (Haybron 2008, 180)

Claudia might be realizing the majority of her most important values to a very high degree, even if *some* of her values must be frustrated, given Haybron's description of the case. The point is not that we should say that Claudia is faring *overall badly*. The point is that she is not faring *especially* well, overall. The values-based theory, though, appears to imply that she *is* doing rather well, because it does not accord much welfare-disvalue to her being unhappy, stressed out, emotionally unfulfilled, short-tempered, anxious, or mildly depressed. Again, something seems amiss.

IV. FROM THE VALUES-BASED THEORY TO WELFARE AS ROBUST AGENTIAL FUNCTIONING

Adjustments can be made to the values-based theory to help address these worries. The first of such adjustment can be called *dispositionalization* (Raibley 2013).

Most theories of personal well-being treat welfare as a purely *actual* property of the agent, so that a person's degree of welfare at a time depends exclusively on what is happening in the actual world-state at that time. Hedonism, for example, states that a person is faring well to the degree that their actual hedonic-doloric balance is positive. On this theory, the welfare-level of a person depends exclusively on their intrinsic, categorical properties in the actual world-state. The person's dispositions are not directly relevant to their welfare. If the person is, e.g., highly irrational and imprudent, and their positive hedonic-doloric balance is due entirely to luck, so that they very nearly had a life full of pain and suffering, this is not bad in itself. Rationality and prudence are only instrumentally valuable; they are good as a means for a person when they produce pleasure, bad as a means when they produce pain, and otherwise indifferent.

The thought that welfare is a purely actual property of the agent seems to be supported by at least the following two lines of thought. First, it is often said that the building blocks of welfare ought to be episodes or events or states that are desirable simply for their own sakes. Pleasure is desirable for its own sake, whereas things like rationality and prudence are normally desired as means. Second, it seems to make sense that, if one is concerned exclusively with one's own welfare,

it makes sense to care only about one's condition in the actual world-state. What happens in 'nearby' but non-actual possible world-states is not something a purely prudentially minded person ought to care about, because they will never experience what happens in those world-states.[3]

These arguments for the 'actualist' take on the welfare-concept are more problematic than they initially appear. First, while the fact that some state is normally desired for its own sake—and the fact that most people would continue to desire this state even if it did not have its customary effects or consequences—might make the state a candidate for being a fundamental bearer of Moorean intrinsic value (Moore 1903), these facts have no clear bearing on whether the state is or is not a bearer of noninstrumental *welfare*-value. Additional premises are necessary to establish this conclusion—e.g., the premise that people normally desire all the intrinsic elements of welfare for their own sake. But this premise is questionable. Even if welfare, itself, is customarily desired for its own sake, this does not guarantee that its component parts are. Furthermore, even if welfare must be something that one would find in some degree compelling or attractive if one were rational and aware (cf. Railton 2003), this does not mean that each of welfare's ingredients must individually be attractive to people who have never thought carefully about welfare's nature. The second argument above is also inconclusive. There is very good reason to care about what happens in 'nearby' but non-actual world states: we frequently do not know which state within a certain range is going to be actual. Perhaps the welfare-concept, itself—at least as it is deployed by laypeople and by medical experts—is shaped by this concern. Just as some argue that the concept of moral rightness has built right into it features that reflect our epistemic limitations—so that what we morally ought to do must be analyzed in terms of what can *rationally be expected* to best promote moral value, as opposed to what actually *will* maximize value—the welfare-concept might also reflect our epistemic limitations. The welfare concept might not suit human cognitive or practical purposes if it was built instead for use by omniscient beings.

Noting these problems with the arguments for the 'actualist' take on the welfare-concept should at least earn dispositionalization a hearing. This is the idea that it makes better sense of ordinary talk involving the welfare vocabulary—and of discourse among experts, including especially medical and psychiatric professionals—to understand welfare as a partly dispositional property. Which dispositions might be directly relevant to welfare? How well an individual is faring at a time (or over a period of time) intuitively depends on the presence of certain 'basic goods' beyond values-realization. For example, it plausibly depends on whether the individual is happy, where happiness is conceived as a diachronic emotional condition that consists in good cheer, engagement, and attunement.[4] It also plausibly depends on other dispositional states, including physical and psychological health; rationality; as well as other character traits and emotional propensities. Given the accepted ways of understanding dispositions (e.g., Lewis 1997), whether

people instantiate these properties depends on how they *would* behave or feel in various non-actual circumstances. Accordingly, welfare depends not just on a person's condition in the actual world-state, but also on their condition in relevant 'nearby' possible world-states.

It seems plausible that the dispositional states just mentioned are not only the dispositions that are intuitively relevant to welfare, but also the components of the broader disposition to realize one's values through one's own activity. It therefore seems that the values-based theory of welfare can be modified in a principled way to accommodate the dispositional nature of welfare. The guiding and unifying thought would be that one's level of welfare depends on whether one is realizing one's values through one's own activity *and* on whether one is stably disposed to do so. Whichever states serve as the causal basis for the disposition to realize one's values will also be components of welfare, itself, so that when these states are present, this counts as directly and not merely instrumentally beneficial. These same states can therefore legitimately be thought of as "basic goods." The diachronic emotional condition of happiness; physical and psychological health; rationality; functionally appropriate emotional states (feeling good when one achieves goals; experiencing negative emotions when one's values are threatened or destroyed); and other personal characteristics (e.g., optimism, good judgment) are *directly* (and not just instrumentally) welfare-constituting, because they underwrite and serve as the causal basis for one's disposition to realize one's values.

The proposal that welfare is partly dispositional has noteworthy implications for the problems discussed in section 3. It can help to address the concerns raised by Arpaly's case of the anorectic. This woman is destroying her physical and emotional health by living in the way that Arpaly describes; the realization of her weight-related values simultaneously amounts to the destruction of dispositions to achieve her other values. With nutritional deficiencies, an impaired immune system, diminished strength, and compromised mobility, it is more difficult for a person to achieve success in their career or to care for loved ones, to play sports or dance or enjoy physical exercise, or even to sustain friendships. Whatever success a person does have in these domains is unstable or lucky if they lack physical and emotional health to a high degree. It is not merely that anorexia nervosa *leads to* hedonic ills, or that it *prevents* the achievement of personal goals: since the states involved in physical health underwrite the disposition to realize one's values, valuing their opposite amounts to identifying with a desire to directly harm oneself. It follows that, even if the anorectic manages to function—even if, by luck, she avoids many of the pains normally associated with her condition—her life cannot be going exceptionally well for her. Indeed, it counts as a subpar life in terms of personal welfare, even if she sincerely judges that she approves of it and is satisfied with it! More generally, it can never, on a dispositionalized values-based approach, be an unmitigated or unqualified benefit to value and achieve pain, frustration, bodily mutilation, humiliation, and the like, because the realization of these values

is directly bad for human beings, even when it is simultaneously directly good *qua* episode of value-realization.

A second adjustment to the values-based theory can be labeled 'holistic appraisal' (Raibley 2012b). The adjustment just discussed, dispositionalization, is already a move in this direction. For dispositionalization requires us to judge, not merely the degree to which an individual is realizing subjectively important values at a time, and not merely the degree to which the individual is stably disposed to do so, but *the degree to which they are doing both these things at once*. Estimating welfare using the dispositionalized theory therefore requires us to make a kind of holistic appraisal of the state of the subject.

However, things are actually even more complicated, and an even more thoroughly holistic form of appraisal is necessary. One way to explain the need for holistic appraisal is to consider the close connection between the activity of values-realization and functioning agency. Now, the term 'agency' is used in many different ways within philosophy, and some of these ways cannot be fully disentangled from broader metaphysical pictures that carry some problematic assumptions. For example, some Kantian ethicists believe that *agents* are beings endowed with reflective capacities that ground the capacity for representing the moral law to oneself and acting out of respect for it. Or, some agent-causal libertarians about free will believe that agents are beings who have the power to act in ways that violate *the causal closure of the physical*, i.e., the principle that no physical events have nonphysical causes. It is not being claimed, here, that welfare is intimately connected with agency in either of these senses. But there is a weaker and more theoretically neutral concept of agency: the capacity rationally to deliberate about what to do, including the capacity to set ends and to initiate and perform complex sequences of action that advance these ends. The connection proposed between welfare and agency is this: since choosing, having, and realizing values *just is* the exercise of agency, if the values-based theory is true, then welfare requires functioning impressively over time as an agent, at least in these respects.

This suggests—though it does not strictly prove—that high welfare requires other aspects of impressive agential functioning. For instance, since we are interested in the welfare of agents that exist and act over such intervals of time, perhaps we should also look at the relations between agents' values as these values evolve over time. For individuals with projects that are synchronically or diachronically incoherent cannot be functioning impressively over time as rational agents; they should not be judged as faring well if the connection between agency and welfare just proposed is indeed a fundamental insight of the values-based theory. Similarly, one will function robustly over time as a rational agent only if one's values are continuous with and responsive to one's experience—i.e., if they are generated by and evolve in response to information about oneself and the world, as opposed to popping into existence without rhyme or reason.

Following this line of thought, estimations of welfare at a time (or of the welfare-values of human lives) would require us to make multifactor comparisons

between actual agents and an exemplar—i.e., a partially specified model or paradigm—of robustly functioning agency. This exemplar has the following features. First, the degree to which an agent is realizing his or her values is a central aspect of robustly functioning agency. So is the degree to which the agent is stably disposed to realize these values—i.e., the degree to which the agent is in good emotional and physical health and has the habits of thought and action mentioned above. Two further aspects of robustly functioning agency were also just identified: the degree to which the agent's values are *coherent*, or jointly realizable over time, and the degree to which they are *responsive to evidence,* i.e., the degree to which they emerge and evolve in response to information about both the external world and the agent's own affective nature, as opposed to emerging and evolving by random accident. Finally, estimations of the welfare-values of human lives would also need to take into account other features, e.g., these lives' durations (Raibley 2012b).

This all suggests the following theory. A person is faring well to the degree that his or her life resembles or matches this model of robust agential functioning; a person is faring badly to the degree that he or she departs from this same paradigm. A particular condition is directly beneficial at a time if its possession by the agent at that time makes for greater resemblance to this model; a particular condition is directly harmful at a time if its possession would constitute departure from this model. A particular condition is overall beneficial if it makes for greater resemblance to the paradigm case of robust agential functioning over a lifetime; a particular condition is overall harmful if it makes for departure from this same model. This theory, which I have sometimes called 'welfare as agential flourishing', is distinct from the values-based theory of welfare, though closely related to it.

These modifications to the values-based approach would indeed allow us to supplement our response to the objection from paradoxical values. It is true that, on the values-based approach, the value that consists in *one's own overall ill-being* is realized just in case it is not realized. This value is therefore paradoxical if the values-based approach is true, though it is not *transparently* paradoxical. However, since the nature of welfare is not transparent, some facts concerning it may not be, either. Furthermore, following up on the suggestion at the end of section 2.3, we can illustrate how an agent might achieve his own overall ill-being, and this will serve to explain away some of the intuitive force of this objection. If an agent were to value pain, frustration, negative emotions, humiliation, and early death—and if he acted consistently to undermine his own physical and psychological health, increasing his own sadness and anxiety and worrying himself into a semi-paralyzed state—then he could truly be said to be realizing his overall ill-being. While he would be successfully realizing such values, and (on the modified values-based theory just described) deriving some benefit from doing so, his overall condition represents such a dramatic departure from the paradigm of robustly functioning agency, that he would count as doing or faring badly. Therefore, the modified values-based theory is compatible with the thought that it is possible to pursue and successfully realize one's own ill-being.

This modified theory also permits us to say that the satisfaction of mere desires is sometimes directly good (this problem was discussed in section 3.1 above). It would have been problematic to simply add that desire-satisfaction is also good, even though value-realization is best. This change would have been somewhat *ad hoc,* and it would have made for a basically disjunctive theory. Furthermore, it would have accorded positive welfare-value to the satisfaction of addictive and compulsive desires and urges. But within the context of the agential flourishing approach, it makes sense to say that the fulfillment of minor, merely desired ends is also beneficial *so long as it does not clash with one's overarching values.* For the fulfillment of desires through one's own activity does constitute successful goal-directed functioning. The satisfaction of such desires is therefore directly beneficial to a small degree, provided that the desires in question are in harmony with the basic values, intentions, and plans of the agent—and provided that the cognitive and physical effort that their satisfaction requires does not impede or detract from success in one's valued endeavors.

This modified theory also allows us to say some additional things about both Lynn (Arpaly's repressed homosexual) and Claudia (Haybron's unhappy attorney). These agents' values are not ideally continuous with or responsive to their experiences in the special sense described above. Lynn "discovers that she is a lesbian" (Arpaly 2003, 16) and that she desires same-sex sexual contact. But she fastidiously avoids it. While the case is in some respects under-described, we can at least say the following. If the case is to support Arpaly's ultimate point, it must be that Lynn feels unfulfilled (if she is celibate) or else frustrated (if she has heterosexual relationships). It is because she feels that "something is missing" that we are inclined to accept that she is a lesbian, as opposed to a bisexual with some repressed homosexual tendencies, or a person with random or alien sexual urges.

This indicates that Lynn's values are not responsive to evidence in several ways. First, she is not modifying her values (e.g., her value of avoiding same-sex sexual contact) when their pursuit brings her no joy or pleasure. In terms of her sexuality, she is locked into what we might call 'prevention-values'—values based on perceived obligations, the achievement of which merely brings quiescence or a feeling of security, as opposed to any positive emotion. When it comes to her sexuality, she lacks 'promotion-values'—values that bring pleasure, joy, and excitement when they are pursued and fulfilled (cf. Higgins, Grant, and Shah 1999). If this is one's dominant orientation within a sphere of life as central and important as one's sexuality, one can hardly be flourishing as an agent. Now, it is possible that Lynn is committed to a religion or ideology that condemns homosexuality, which might complicate the case in various ways. But if she accepts a secular, scientific worldview, and mainly fears the disapproval of friends and family members, then her values are also not responsive to empirical evidence. For example, they are not responsive to the fact that good nonreligious arguments for the immorality of homosexual behavior are difficult (if not impossible) to come by; or the fact that many people live happy, healthy, and fulfilled lives as homosexuals; or the fact that

there are plenty of people who would be perfectly accepting and supportive of her choosing a lesbian lifestyle. And so when it comes to the continuity of her values with her experience, she departs from the paradigm, and she is therefore faring less well than she might be.

Similarly, Claudia pursues wealth and status, which apparently brings her periodic bursts of positive affect, but which leaves her short-tempered, stressed-out, anxious, and depressed. While she might be happy in the episodic sense briefly described at the outset, she is not happy in the emotional condition sense described by Haybron. This case may also be under-described in significant ways. There must be some reason that Claudia pursues these values and feels this way. One way of filling out the case is that Claudia's valuing of wealth and status is at root *inauthentic*. Perhaps her parents implanted in her loyalty to these ideals by criticizing her and withholding affection when she acted for the sake of other values; perhaps they filled her head with bad arguments for the objective value of wealth and social status that she is still ill-equipped to criticize. The problem with this way of filling out the case is that the achievement of imposed ideals of this kind usually *does* bring a kind of quiescence or tranquillity, at least according to a leading model of human motivation (Higgins 2012). The typical problem with such lives is that they are experienced as *joyless* and *empty,* not as anxiety-ridden and stress-filled. And so, holding constant facts about human motivation, the interpretation of the case that fits better with Haybron's description is that these are Claudia's *own* values—they are really what she cares about—but that the work of lawyering is just ill-suited to her emotional nature.

If this is indeed Claudia's problem, then welfare as agential flourishing can say something helpful about the case: Claudia's values are not responsive to evidence, in that she does not modify them on account of the fact that their pursuit in real life is accompanied by burdensome and painful emotions. In this respect, she is not functioning robustly as an agent.

However, this reply might not go far enough. Even though Claudia fails to adjust her values in response to negative affective feedback in the way characteristic of maximally efficacious human agents, she might still count as quite stably disposed to realize her values. Perhaps her intense craving for wealth and social status stems from being impoverished and powerless when she was very young. This craving might have persistent motivational force. She might also have found ways to manage the negative emotions she experiences while pursuing her goals, so that she functions at a high level in spite of these emotions' general tendency to be deflating. Simultaneously, Claudia's values might be highly responsive to evidence in other ways; they might evolve appropriately in response to new information about things other than her own affective and motivational nature, such as changing market conditions and technological developments. For all these reasons, she may not be so far from the paradigm of robustly functioning agency. And so it seems that welfare as agential flourishing implies that she is faring well, even if not superlatively well. But this result seems unacceptable.

V. HIDDEN ASPECTS OF AGENCY

In several of the problematic cases we have discussed—Lynn, Arpaly's anorectic, and Claudia—there appears to be a mismatch or clash between lower-order affective, appetitive, and drive-states (on the one hand) and the subject's consciously held values (on the other). Though she consciously values them, Lynn does not really *like* or *enjoy* celibacy or heterosexual relationships in the way she would like and enjoy homosexual relationships. The anorectic's consciously held values are at odds with her drive to alleviate pangs of hunger. Claudia achieves goals the pursuit of which is accompanied by burdensome emotions. It is tempting to say that incoherence among one's lower-order affective, appetitive, and drive-states is (other things being equal) directly harmful, as is incoherence between these states (on the one hand) and one's consciously held values (on the other). However, it might be thought that this proposal is basically at odds with the values-based theory's fundamental rationale, because it implies that values sometimes ought to be modified so that they cohere with lower-order states. In this section, it will be argued that this modification is in fact available to values-based theorists—and indeed, it is a natural extension of the theory—provided that they embrace the idea (stated in the previous section) that values-realization is important for welfare *because* it is a central aspect of functioning agency.

We should begin by asking, "Why were values invoked by welfare-theorists in the first place?" One possible answer is that values require some procedure of reflective endorsement that will, simply in-and-of-itself, sort the pro-attitudes whose satisfaction would benefit an agent from those whose satisfaction would not. But as Hilary Kornblith points out, on all the obvious ways of specifying the relevant procedures, there is the real possibility that someone could run through the relevant procedures and use them to rationalize pro-attitudes that are problematic. He writes: "There is no second-order magic. Second-order mental states are not so very different from first-order mental states: both are firmly entwined in the same causal net; both are, at times, reasons-responsive, and, at times, disengaged from reason" (Kornblith 2010, 18–19). Whatever errors can be made at the level of forming ordinary pro-attitudes like beliefs and desires can also be made at the level of forming higher-order pro-attitudes such as values.

But there is another possible answer to the question, "Why were values invoked in the first place?" It is this. By looking at values (pro-attitudes that are adequately informed, stable, identified with by the subject, and treated as justificatory of action), we will be focusing on pro-attitudes that (a) are responsive to lots of important information that has been processed in sophisticated ways, (b) are consciously accessible, so that they can enter into planning and deliberation, and (c) are typically coherent with the agent's larger affective and motivational psychology, on account of which they can motivate and sustain purposeful activity, especially complex sequences of action. In having these features, values make

rational agency possible. In virtue of their cognitive aspects, values provide input for complex, volitionally guided practical reasoning that is responsive to information. Simultaneously, in virtue of their affective aspects, values sustain activity over the long haul and in the face of adversity, even if they may not always 'win out' when they are pitted against strong lower-order pro-attitudes. It is for *these* reasons that an agent's degree of value-realization is often a reliable proxy for an agent's degree of personal welfare. This is also why an agent's degree of desire-satisfaction (where this includes the fulfillment of urges, compulsions, and whims) is a less reliable indication of the agent's welfare.

This implies that values are usually—but not invariably—authoritative when it comes to what would benefit a person. It also suggests that values may *fail* to be authoritative precisely in those cases where they cannot be made to cohere with certain fixed aspects of the agent's affective or motivational psychology. For it would be a mistake to suggest that *all* action ought to be governed exclusively by consciously held values and beliefs—or that values automatically 'trump' nonconscious or subpersonal drives, states, and processes when it comes to what would benefit the agent. Such drives, states, and processes, alongside lower-order desires and affective states, also support our ability to act. Most obviously, some drives spur action that satisfies basic needs. Additionally, according to a model of the mind favored by leading cognitive scientists and neuroethicists (Baars 1997; Dehaene and Naccache 2001; Wilson 2002; Levy 2007), consciously held values 'broadcast' information so that such subpersonal motivational and cognitive systems can "go to work" on them, which serves to unify our thought and behavior so that we avoid self-defeat. If this is true, there is no reason to think of these as lesser aspects of the self to be repressed, overcome, or eliminated by a purer, rational self. We would be unable to successfully cope with the world without them. Fruitful agency involves respecting and utilizing these aspects of the self.

Agency would be impossible without values, but it would also be impossible without lower-order drives, attitudes, affective reactions, and cognitive processes. When we think of the paradigm case of the flourishing agent, we should therefore think of a person with a harmoniously functioning cognitive and motivational system—i.e., a person with minimal conflict among his or her lower-order states and processes, and minimal conflict between these and higher-order states and processes. We must therefore add to the model of robustly functioning agency introduced in section 4. The optimally functioning agent has and realizes a robust array of values. This agent is also stably disposed to realize these values, on account of being in good physical and psychological health, having functionally appropriate affective responses (including, in some cases, *negative* affect), and possessing various character traits. This agent's values are *coherent*, or jointly realizable over time, and *responsive to evidence,* including information about the external world and the agent's own affective nature. Finally, there is minimal conflict among the agent's lower-order drives, attitudes, and states—as well as minimal conflict between these states and higher-order ones. Resembling this paradigm at a time

makes for a high degree of personal welfare, and a life that goes well for the one who lives it resembles this paradigm over a normal human life span.

This approach builds on insights from other theorists. Both Dan Haybron (2008) and Valerie Tiberius and Alexandra Plakias (2010) emphasize the importance of affective nature. Haybron holds that a person's emotional or affective nature must also be brought to fruition (i.e., developed or fulfilled) in order for the person to be doing well. For example, it would benefit Claudia (in the case above) to recognize that the pursuit of these values leaves her chronically unhappy and emotionally burdened and make appropriate changes in her goals. However, it is inexact to say, as Haybron does, that Claudia ought to 'fulfill' her emotional or affective nature (Raibley 2012a). Rather, as our modified theory implies, her consciously held values ought to be made coherent with her affective nature.

Similarly, Tiberius and Plakias write that an agent's activities, relations, goals, and so forth are not fully their values if these items are ill-suited to the agent's affective nature, as this greatly undermines the *stability* of one's identification with them. But, given the details of these cases, it is not clear that the values professed by Lynn and Arpaly's anorectic really are inherently unstable. Nor should we be content to say that the realization of these values is directly beneficial when it occurs, but overall bad. The real problem, as the modified theory would have it, is that Lynn and the anorectic's consciously held valuations are *in conflict with subpersonal drives and affective dispositions that cannot easily be changed.* This incoherence is directly and intrinsically harmful.

By locating a deeper rationale for the values-based approach in the concept of robustly functioning agency—and by looking to contemporary medicine, clinical psychology, cognitive science, and neuroethics to understand the elements of such agency—well-being as agential flourishing can explain why these individuals' consciously held evaluations ought to be changed to cohere with lower-order attitudes. Fully functioning agency not only requires value-realization, bodily and emotional health, functionally appropriate affect, and coherent and evidence-responsive values; it also requires congruence among values, subpersonal drive-states, and affective dispositions. Simultaneously, this modified theory helps to answer traditional objections to the values-based approach based on self-sacrifice and disinterested and paradoxical values. For these reasons, it is worthy of further attention and development.

ACKNOWLEDGMENTS

Special thanks to Guy Fletcher for comments on an earlier draft; to Chris Heathwood and Valerie Tiberius for discussion of the material in section 2; and to Teresa Chandler for discussion of the objections in section 3. Thanks also to Anna Alexandrova, Erik Angner, Donald Bruckner, Havi Carel, Dan Hausman, Dan

Haybron, Antti Kauppinen, Christopher Megone, Peter Railton, Sam Wren-Lewis, and members of the audience at the University of Leeds conference, Measures of Subjective Well-being for Public Policy, for helpful comments and discussion.

NOTES

1. Raibley (2013) contains a discussion of cases that involve agents with multiple desires, including some desires for self-harm.
2. Thanks to Donald Bruckner for helpful discussion of this objection.
3. These points have been illustrated using hedonism, but similar things are true of desire-satisfactionism. Though some desire-satisfactions employ a motivational and dispositional conception of desire, and so may look at "nearby" world-states to determine which desires the agent has, whether these desires are satisfied depends only on what is happening in the actual world-state. Similarly, life-satisfactionists look exclusively at whether the individual judges that he or she is actually satisfied with the life he or she is actually leading. Even leading versions of the Objective List Theory are largely 'actualist', though states like knowledge may turn out, on further analysis, to involve dispositions.
4. According to Haybron, attunement involves self-confidence, clarity of mind, and an open (as opposed to defensive) posture. It is manifest by those who are "comfortable in their own skin," and by those who are not preoccupied with defense reactions to threats coming from outside. Roughly speaking, it is the opposite of being stressed (Haybron 2008).

REFERENCES

Adams, R. M. 1999. *Finite and Infinite Goods*. New York: Oxford University Press.
Anderson, E. 1993. *Value in Ethics and Economics.* Cambridge, MA: Harvard University Press.
Arpaly, N. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. New York: Oxford University Press.
Baars, B. J. 1997. *In the Theater of Consciousness*. New York: Oxford University Press.
Boyd, R. 1988. "How to Be a Moral Realist." In *Essays on Moral Realism*, ed. G. Sayre-McCord. Ithaca, NY: Cornell University Press.
Brandt, R. 1979. *A Theory of the Right and the Good.* Oxford: Clarendon Press.
Copp, D. 1995. *Morality, Normativity, and Society.* New York: Oxford University Press.
Crisp, R. 2005. *Reasons and the Good.* New York: Oxford University Press.
Dehaene, S., and L. Naccache. 2001. "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework." *Cognition* 79: 1–37.
Diener, E., R. A. Emmons, R. J. Larsen, and S. Griffin. 1985. "The Satisfaction with Life Scale." *Journal of Personality Assessment* 49: 71–75.
Diener, E., and R. Biswas-Diener. 2008. *Happiness: Unlocking the Mysteries of Psychological Wealth.* New York: Wiley-Blackwell.
Feldman, F. 2004. *Pleasure and the Good Life.* New York: Oxford University Press.
Feldman, F. 2010. *What Is This Thing Called Happiness?* New York: Oxford University Press.
Frankfurt, H. 1988. *The Importance of What We Care About.* New York: Cambridge University Press.
Gaus, G. 1990. *Value and Justification.* New York: Cambridge University Press.
Goldman, A. 2009. *Reasons from Within*. New York: Oxford University Press.
Griffin, J. 1986. *Well-being*. Oxford: Clarendon Press.
Harman, E. 2009. "Harming as Causing Harm." In *Harming Future Persons,* ed. M. Roberts and D. Wasserman, 137–54. Dordrecht, Netherlands: Springer.
Haybron, D. 2008. *The Pursuit of Unhappiness*. New York: Oxford University Press.

Haybron, D., and V. Tiberius. 2012. "Normative Foundations for Well-being Policy." In *Papers on Economics and Evolution,* ed. Evolutionary Economics Group. Jena, Germany: Max Planck Institute.
Heathwood, C. 2005. "The Problem of Defective Desires." *Australasian Journal of Philosophy* 83: 487–504.
Heathwood, C. 2011. "Preferentism and Self-Sacrifice." *Pacific Philosophical Quarterly* 92: 18–38.
Higgins, E. T., H. Grant, and J. Shah. 1999. "Self-Regulation and Quality of Life: Emotional and Non-emotional Life Experiences." In *Well-being: The Foundations of Hedonic Psychology,* ed. D. Kahneman, E. Diener, and N. Schwarz 244–66. New York: Russell Sage Foundation.
Higgins, E. T. 2012. *Beyond Pleasure and Pain: How Motivation Works.* New York: Oxford University Press.
Hubin, D. C. 1999. "What's Special about Humeanism." *Noûs* 33: 30–45.
Hubin, D. C. 2003. "Desires, Whims, and Values." *Journal of Ethics* 7: 315–35.
Kagan, S. 1994. "Me and My Life." *Proceedings of the Aristotelian Society* 94: 309–24.
Keller, S. 2004. "Welfare and the Achievement of Goals." *Philosophical Studies* 121: 27–41.
Keller, S. 2009. "Welfare as Success." *Noûs* 43: 656–83.
Kornblith, H. 2002. *Knowledge and Its Place in Nature.* New York: Oxford University Press.
Kornblith, H. 2012. *On Reflection*. New York: Oxford University Press.
Korsgaard, C. 1996. *The Sources of Normativity.* New York: Cambridge University Press.
Kraut, R. 1994. "Desire and the Human Good." *Proceedings and Addresses of the American Philosophical Association* 68: 39–54.
Levy, N. 2007. *Neuroethics*. New York: Cambridge University Press.
Lewis, D. 1997. "Finkish Dispositions." *Philosophical Quarterly* 47: 143–58.
Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. New York: Penguin Books.
Moore, G. E. 1903. *Principia Ethica.* New York: Cambridge University Press.
Nagel, T. 1979. *Mortal Questions.* New York: Cambridge University Press.
Nozick, R. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.
Parfit, D. 1984. *Reasons and Persons*. Oxford: Clarendon Press.
Portmore, D. 2007. "Welfare, Achievement, and Self-Sacrifice." *Journal of Ethics & Social Philosophy* 2: 2, www.jesp.org.
Raibley, J. 2010. "Well-being and the Priority of Values." *Social Theory and Practice* 36: 593–620.
Raibley, J. 2012a. "Happiness Is Not Well-being." *Journal of Happiness Studies,* 13 (6): 1105–29.
Raibley, J. 2012b. "Welfare over Time and the Case for Holism." *Philosophical Papers* 41: 239–65.
Raibley, J. 2013. "Health and Well-being." *Philosophical Studies* 165: 469–89.
Railton, P. 2003. *Facts, Values, and Norms.* New York: Cambridge University Press.
Rawls, J. 1971. *Theory of Justice*. Cambridge, MA: Belknap Press.
Rice, C. 2013. "Defending the Objective List Theory of Well-being." *Ratio* 26: 196–211.
Scanlon, T. M. 1998. *What We Owe to Each Other.* Cambridge, MA: Belknap Press.
Schwarz, N., and F. Strack. 1999. "Reports of Subjective Well-being: Judgmental Processes and Their Methodological Implications." In *Well-being: The Foundations of Hedonic Psychology,* ed. D. Kahneman, E. Diener, and N. Schwarz. New York: Russell Sage.
Shiffrin, S. 1999. "Wrongful Life, Procreative Responsibility, and the Significance of Harm." *Legal Theory* 5: 117–48.
Sumner, L. W. 1996. *Welfare, Happiness, and Morality*. Oxford: Clarendon Press.
Tiberius, V. 2000. "Humean Heroism: Value Commitments and the Source of Normativity." *Pacific Philosophical Quarterly* 81: 426–46.
Tiberius, V. 2008. *The Reflective Life.* New York: Oxford University Press.
Tiberius, V., and A. Hall. 2010. "Normative Theory and Psychological Research: Hedonism, Eudaimonism, and Why It Matters." *Journal of Positive Psychology* 5: 212–25.
Tiberius, V., and A. Plakias. 2010. "Well-being." In *The Moral Psychology Handbook,* ed. J. Doris et al. New York: Oxford University Press.
Wilson, T. 2002: *Strangers to Ourselves: Discovering the Adaptive Unconscious.* Cambridge, MA: Belknap Press.