Philos Stud (2016) 173:1203–1232
DOI 10.1007/s11098-015-0527-9

Self-expression: a deep self theory of moral responsibility

Chandra Sripada[1]

Published online: 12 August 2015

Abstract According to Dewey, we are responsible for our conduct because it is "ourselves objectified in action". This idea lies at the heart of an increasingly influential *deep self* approach to moral responsibility. Existing formulations of deep self views have two major problems: They are often underspecified (for example, they rely heavily on metaphorical language), and they tend to understand the nature of the deep self in excessively rationalistic terms. Here I propose a new deep self theory of moral responsibility called the Self-Expression account that addresses these issues. The account is composed of two parts. The first part answers the question, What is a deep self? Theorists have tended to favor cognitive views that understand the deep self in terms of rationally formed evaluative judgment. I propose instead a conative view that says one's deep self consists of a distinctive kind of pro-attitude, cares, and I provide an account of cares in terms of their distinctive psychological functional role. The second part answers the question, When does an action express one's deep self? I criticize the agentially demanding conditions set out in existing views and propose a more minimalist alternative. I show that the Self-Expression account handles issues that bedeviled traditional deep self views, including how to explain moral responsibility for spontaneous, out of character, and weak-willed actions.

Keywords Moral responsibility · Deep self · Real self · Expression · Endorsement · Susan Wolf · Gary Watson · Harry Frankfurt

✉ Chandra Sripada
sripada@umich.edu

[1] Department of Philosophy, University of Michigan, 2215 Angell Hall, 435 South State Street, Ann Arbor, MI 48109-1003, USA

1 Introduction

One influential approach to moral responsibility focuses on the sources of a person's actions. This approach distinguishes actions that originate in factors internal to the person versus those that are external to the person. Actions that have the appropriate sort of internal source *belong* to the person in a distinctive sort of way that renders her morally responsible for them.

A challenge for this approach to moral responsibility is to formulate the correct criterion of internal sourcehood. One simple and historically influential view says this criterion is met whenever a person's actions arise from her strongest desires. On this view, I would be morally responsible for pushing over a man if the pushing was what I most wanted to do, but not if I am propelled into the man against my will.

A problem with this view, however, is that it appears to set too low a bar. Harry Frankfurt asks us to consider the case of an addict whose strongest desire is to use a narcotic, but the man is profoundly alienated from this desire.

> [The man] hates his addiction and always struggles desperately, although to no avail, against its thrust. He tries everything that he thinks might enable him to overcome his desires for the drug. But these desires are too powerful for him to withstand, and invariably, in the end, they conquer him. He is an unwilling addict, helplessly violated by his own desires (Frankfurt 1971, p. 12).

Even though using the narcotic is the man's strongest desire, it is plausible that he is not morally responsible for using the drug.[1] This case—and related cases involving kleptomaniacs, compulsives, and victims of brainwashing—suggest the need to formulate a more refined criterion for internal sourcehood, one that recognizes that desires that lie within the physical boundaries of the person's psychology may nonetheless be external to the person's self.

Deep self theories of moral responsibility are attempts to do just this. A variety of deep self theories have been put forward, and these theories differ in important ways. Nonetheless, certain broad commonalities can be discerned. All deep self theories share the view that, of the totality of attitudes in a person's psychology, there is a distinguished subset of them that are fundamental to her practical identity. These attitudes, dubbed the "deep self" or "self" for short, belong to her in a distinctive way that carries significance for a number of aspects of agency, and most centrally for our purposes, moral responsibility. Now, the criteria used to pick out this subset of deep attitudes vary from theory to theory. But once this subset is

[1] Assume that the addict did not knowingly get himself addicted in the first place. Without this stipulation, then the case concerns not *direct* responsibility, but rather the complex issue of *derived* responsibility. This latter sense of responsibility applies when an agent *A*'s, some prima facie responsibility-undermining factor *F* is present, and the agent in some appropriate way (e.g., responsibly, knowingly, and intentionally) brought *F* about. When these conditions obtain, the agent is said to be responsible *in a derived sense* for *A*-ing. Unless otherwise explicitly noted, it is direct responsibility, rather than derived responsibility, that will be my exclusive focus throughout this article.

specified, all deep self views agree that a person is morally responsible for an action only if it expresses her deep self.[2]

Deep self views, unlike the desire-based view mentioned earlier, *can* make sense of why Frankfurt's Unwilling Addict is not morally responsible for using the drug. Though the desire to use the narcotic is indeed his strongest, this desire does not reflect, and indeed dramatically opposes, his deep self, and thus he is not responsible for the actions that issue from this desire.

Deep self theories are rapidly gaining in popularity. Treatments of moral responsibility that strike deep self themes have recently been offered by T. M. Scanlon, Nomy Arpaly, Angela Smith, George Sher, and Sarah Buss, among others.[3] Susan Wolf provided a classic statement of the deep self view and identified precursors of the view in the works of Harry Frankfurt and Gary Watson.[4] P. F. Strawson suggested that our practice of holding others morally responsible is intimately bound up with assessment of the *quality of will* displayed in their actions.[5] Depending on how this idea is filled in, it might very well be understood along deep self lines. More distantly, deep self themes are scattered in the writings of such notables as John Dewey, David Hume, and Aristotle.[6]

Despite enjoying an extensive and eminent pedigree, deep self theories of moral responsibility confront two critical problems. First, there appears to be a good deal of confusion about what these approaches are actually committed to, and this is often due to the fact that theorists rely heavily on metaphors or evocative images. For example, theorists speak about one's self being "objectified in actions", or actions that "flow from" or "disclose" one's deepest self. One aim of this paper is to provide a detailed specification of a deep self theory in which these vivid but

[2] This way of formulating a deep self theory makes expression of the self necessary for moral responsibility but not sufficient—additional criteria must also be satisfied. A standard view is that there is also an *epistemic requirement* for moral responsibility. For example, if a person is ignorant that there is a kitten snoozing behind his car, and if he is completely non-culpable for his state of ignorance, then if he unfortunately backs his car over the kitten, then he is not morally responsible for the kitten's death (see Fischer and Ravizza 1998, p. 13). I agree with the standard view that moral responsibility requires meeting certain epistemic conditions. My position, however, is that these epistemic conditions are already fully *built into* a deep self theory's requirement of self-expression. That is, an agent's *A*-ing can't express himself unless the agent satisfies certain epistemic requirements with respect to his *A*-ing. Thus, there isn't a separate freestanding epistemic requirement that one must meet over and above the requirement of self-expression. This is a large and complex topic, and so I develop this position in detail elsewhere.

[3] See Scanlon (1998), Arpaly (2003), Smith (2005, 2008), Sher (2009), and Buss (2012). Of note, while Scanlon, Arpaly, and Smith use language that is superficially consonant with the deep self approach, and others often interpret them along these lines, I believe their views in fact resist easy categorization. Among other things, these theorists don't draw a deep versus surface distinction, and I believe that any full-fledged deep self view requires this distinction.

[4] See Wolf (1993), Frankfurt (1971), and Watson (1975).

[5] See Strawson (1962).

[6] Dewey discusses the connection between moral responsibility and one's self, understood as one's conception of the good, in Dewey (1957), especially chapter 3. Hume discusses the relationship between responsibility and character in Treatise, bk. 11, Pt. 111, sec. 2. In Nicomachean Ethics, Magna Moralia, and Eudemian Ethics, Aristotle proposes slightly different versions of a theory that morally responsible agency requires a cause that is internal to the agent and expresses his character. See Meyer (2011), especially chapter 4, for a helpful discussion.

ultimately metaphorical expressions are made more precise. A second problem with deep self theories is that they are usually formulated in ways that are, in a sense to be clarified shortly, excessively *rationalistic*. As a consequence, these theories have problems explaining moral responsibility in a host of areas, including spontaneous, out of character, and weak-willed actions. Thus, another aim of this paper is to offer a new deep self theory that is less reliant on highly demanding forms of reasoning and reflection and is instead anchored in our desiderative natures.

The *Self-Expression* account that I propose lies at the intersection of two decision points that any deep self theory of moral responsibility confronts. The first concerns the question of what counts as a person's deep self. Many theorists have offered *cognitive* views in which the deep self consists of certain judgments about what one has most reason to do. In part 1, I propose a *conative* view that says that a person's deep self consists of her *cares*. Cares are a distinctive kind of mental state set apart from other mental states (e.g., beliefs and ordinary desires) by their unique functional role. The distinctive syndrome of dispositions associated with cares, I argue, makes them well suited for constituting a person's deep self. The other decision point concerns what it means to say an action or attitude *expresses* a person's deep self. In part 2, I criticize views that place highly agentially demanding conditions—such as reflective endorsement or governance by one's valuational system—on the expression relation and propose a more minimalist alternative. In the remainder of part 2 and in part 3, I show that the Self-Expression account provides an attractive explanation for how agents can be morally responsible for diverse kinds of actions and conduct. This includes spontaneous emotions, non-deliberative conduct, actions performed outside of conscious awareness, actions performed by agents lacking well-developed reflective capacities, conduct that is out of character, and weak-willed actions.

2 What is a deep self?

The first question any deep self theory of moral responsibility has to answer is what attitudes constitute a person's deep self. I begin by discussing cognitive approaches to the deep self before turning to the alternative that I favor, a version of a conative approach.

2.1 Cognitive approaches to the deep self

In a widely discussed work,[7] Gary Watson distinguishes a person's motivational and valuational systems. The motivational system consists of the set of "non-rational" psychological attitudes that move her to act. These might include appetitive desires, urges, and spontaneous emotions. For some animals, the motivational system might be all there is for the production of action. Human beings, however, also have a valuational system. This system operates on considerations, that, when combined

[7] Watson (1975).

with one's factual beliefs, yield judgments of form: "the thing for me to do in these circumstances, all things considered, is *a*." Watson notes that these two systems can diverge; what a person actually desires and what a person judges to be desirable can be in stark opposition.

Based on Watson's division of the psyche,[8] one might propose an account of the deep self along the following lines: The deep self consists of the judgments issued by one's valuational system (or more specifically, judgments pertaining to matters of significance such as the ways one should lead one's life—I leave this qualifier off going forward). The contents of the deep self specify what is *in fact* genuinely important to the person and most worth pursuing. At least at first pass, it does seem plausible that the reflective verdicts of one's valuational system, construed as separate from the operations of mere desire or appetite, are well suited to play this role.

It can't be, however, that *all* of one's reflectively formed evaluative judgments articulate the perspective of one's deep self. We are all familiar with cases in which, due to various distorting factors, what a person judges to be most worth pursuing comes apart from what in fact she should most pursue. To draw (very loosely) on one of Watson's examples, a man raised under puritanical strictures may judge that pleasures of the flesh are forbidden and to be avoided at all costs (*ibid*, p. 210). This evaluative judgment reflects the man's acculturation, we think. It does not reflect the point of view of his self.

This case highlights an issue that we might term the "Which Judgments?" problem that arises not only for Watson's view, but for any cognitive view that says one's deep self consists of a certain class of evaluative judgments. Among the full set of a person's evaluative judgments made across various times and circumstances, certain judgments genuinely reflect her fundamental practical stance, while certain judgments, such as the one made by the man with puritanical acculturation, do not. A cognitive conception of the deep self must provide some criterion to separate the two.

Watson adds the qualification that the relevant evaluative judgments must be made in "cool and non-deceptive moment[s]" (*ibid*, p. 215). This suggestion is helpful but in the end appears insufficient. To be sure, cool and calm conditions often exert salubrious effects on reasoning, but they certainly do not *ensure* that the judgments that ensue genuinely stand for one's self. The man with puritanical acculturation might under even the most sober conditions continue to judge that sex is sinful (indeed, perhaps it is not calm, cool reasoning but rather raw passion—a state of unbridled lust—that would get the man to finally realize where he truly stands on the issue). Watson's second qualifier, "non-deceptive", is harder to interpret. Under the most plausible reading, the qualifier just means that the judgment *genuinely* reflects one's self. If that is what it is supposed to mean, then no

[8] Watson's model of the psyche is actually more nuanced (for example, it incorporates a role for *values* in addition to the activities of the valuation system). My interest is less in his specific model, but rather the general strategy of dividing the psyche in terms of motivational and valuational activities as a way to understand the deep self.

progress has been made. The criterion for when a judgment genuinely reflects one's self in this way is precisely what was in need of clarification.

Even if the qualifiers "cool and non-deceptive" don't quite do the trick, perhaps there is nonetheless *some* way to specify a more favorable epistemic position that guarantees that judgments made from that position do in fact articulate the perspective of one's self. It is unlikely, however, that this strategy can succeed. It would seem that whatever epistemic circumstances are specified (e.g., the person has no false beliefs, does not make mistakes of reasoning, entertains considerations sufficiently vividly, etc.), it would still be possible to construct clear counterexamples in which the person makes some particular evaluative judgment and this judgment fails to articulate the perspective of his self. Indeed, the problem here bears strong resemblance to an analogous one that arises for ideal attitude theories of what is good for a person, i.e., a person's non-moral good.[9] These theories claim that a person's good is constituted by what she would want (or what she would want herself to want) under idealized epistemic conditions. While these theories have certain attractive aspects, they nonetheless have been faced with a slew of compelling, if not decisive, counter-examples.[10] I suspect that attempts to address the Which Judgments? problem via an idealizing strategy would face similar troubles.

The Which Judgments? problem is an important challenge for cognitive conceptions of the deep self. Later on (Sect. 4.3), I identify additional serious problems. For now, I turn instead to another family of approaches that say one's deep self is constituted not by certain *judgments*, but rather by certain *pro-attitudes*.

2.2 The care-based account of the deep self and the functional role account of cares

I propose a conative account that says one's deep self is constituted by a distinctive class of pro-attitudes, *cares*. People have a dizzying array of desires and other motivational attitudes directed at doing various things: going to the park, playing games with friends, taking vitamins, buying a new computer, and so on. When we look closely at the structure of a person's economy of motivational attitudes, we find that many of these attitudes are arranged in an *instrumental hierarchy*; that is, the person desires to do X only in order to fulfill her desire to do Y, which in turn is in the service of fulfilling desire Z. Consider Katya, who wants to get on the bus. She does this only because she wants to get to class, and this too is done in the service of a series of further desires: fulfilling the organic chemistry requirement, getting her medical degree, becoming a competent physician. When we trace sequences such as these to see where they lead, we often encounter at their very root a distinct class of conative states: *cares*. Katya wants to be a competent physician because she *cares* about helping those who are in need—she wants to relieve their suffering. It seems

[9] See Brandt (1998) and Railton (1986a, b) for classic defenses of this kind of view.

[10] See, for example, Rosati (1995) and Velleman (1988).

intuitively plausible that what a person cares about and her deep self are importantly connected. I want to explore this idea further.

Start by considering the question of what distinguishes cares from other mental states such as beliefs and ordinary desires. I propose that the answer is, in brief, *functional role*. Beliefs are associated with a characteristic functional role that, depending on one's favored account, might include such things as being inferentially promiscuous, exhibiting a mind-to-world direction of fit, and serving as the maps by which we steer. According to my proposed account, cares too exhibit a characteristic functional role. They exhibit a syndrome of dispositional effects that includes motivational, commitmental, evaluative, and affective elements, and I want to spend some time discussing each of these in turn.

Begin with the *motivational* effects of caring. As the example of Katya suggests, caring about something always serves as a source of *intrinsic* motivation for actions that promote the achievement of that thing. It is not possible to care for something *only* to achieve some further end. Notice that when we consider the vast array of motivational attitudes within a person's psychology, since caring requires being a source of intrinsic motivation, it must be fairly rare. Most every other motivational state rises and falls in the service of other more basic motives. Cares are distinctive in lying exclusively at the foundations of this hierarchy of motives.

Second, caring is associated with a distinctive set of *commitmental* effects. In caring about X, the person not only is intrinsically motivated to bring about X, the person is *in addition* intrinsically motivated to continue caring about X.[11] The commitmental features of caring can be brought out more vividly by considering a person's responses to the prospect of changes to the elements of her conative set. If a person's pro-attitude towards X is merely a desire, and in particular a desire that is not in any way instrumental to anything she cares about, then she should be relatively indifferent to the prospect of this attitude being altered in some way: for example, being replaced by some attitude Y (where Y too is similarly irrelevant to her cares). In contrast, if she cares for X, then it would strike us as strange if she were indifferent to the prospect of change—if offered a pill that would erase one of her cares, she says, "Meh, doesn't matter to me if I take this pill. Either way." This does not imply that she can't ever change her cares. Rather, the point is that even if she initiates a change in her cares—for example, she judges that a certain care of hers must be, all things considered, altered or erased—if the attitude at issue is genuinely a care, then the prospect of changing it will not be viewed *only* positively. It must be true that there is at least a small part of her, i.e., that part of her that cares for X, that wants to continue to go on so caring. Thus whether or not it is acknowledged or overtly expressed, altering or erasing one of one's cares will tend to be accompanied by an experience of *loss*.

[11] In his later writings, Frankfurt arrived at a similar view that caring involves commitmental higher-order motivation:

> When we care about something, we go beyond wanting it. We want to go on wanting it, until at least the goal has been reached. Thus, we feel it as a lapse on our part if we neglect the desire, and we are disposed to take steps to refresh the desire if it should tend to fade. The caring entails, in other words, a commitment to the desire (Frankfurt 2006, pp. 18–19).

Third, caring is connected to evaluative judgment; a person who cares about something is disposed to form judgments about that thing that cast it in a normatively favorable light.[12] For example, a person who cares deeply about cultivating his talents in opera is disposed to judge that achieving such talent is good, that it is valuable, or that there is a reason to pursue those actions that promote it.[13] This connection between cares and evaluative judgment is one of the important ways in which cares achieve efficacy in motivating action. During practical reasoning, a person weighs the reasons that favor various alternative courses of action. Moreover, under the right circumstances, one's practical judgments about what to do can motivate action. Since cares dispose a person to judge that there is a reason to perform those actions that promote the satisfaction of the care, these judgments about reasons can, via their role in practical reasoning, eventuate in care-promoting actions.

Finally, caring is also associated with a rich and distinctive profile of emotional responses that are finely tuned to the fortunes of the thing that is the object of the care.[14] These emotions help to shape and refine one's overall suite of cognitive and actional responses. Suppose Paul cares about the plight of children dispossessed by war in Sudan. If a hostile United Nations resolution on Sudan is forthcoming, Paul is disposed to a suite of "signaling" emotions such as anxiety and fear that concentrate his attention on the looming threat and give it precedence over other considerations. If Sudanese children are benefited or advanced in some way, Paul is disposed to a suite of positively valenced emotions such as joy, approval, and elevation. If the fortunes of the Sudanese children are set back, Paul is susceptible to sadness,

[12] The cognitive view of the deep self says that the deep self is *constituted* by a certain class of evaluative judgments. The care-based conative view currently under discussion says that the deep self consists of a class of pro-attitudes, namely cares, which exhibit a syndrome of effects, *one* of which is to *dispose* the person to form care-concordant evaluative judgments. The two views might seem superficially similar, but there are a number of ways in which they in fact differ. First, the cognitive view places *all* elements of the relevant class of evaluative judgments within one's deep self. The care-based view allows that many evaluative judgments don't bear any connection to the deep self, namely those that don't bear the right dispositional tie to one's cares. Second, notice that dispositions might not ever be realized—they can be *defeated* by a mask or fink or antidote. Thus the care-based view is consistent with a person's caring for X and, due to the operation of a defeater, failing to make the relevant evaluative judgments regarding X, even in idealized epistemic circumstances (see also footnote 17). A third respect in which the two views differ is in how they handle conflict within one's deep self. I take up this idea later (see Sect. 4.3).

[13] There is a further claim that one might make: The truth of these evaluative judgments—that is, judgments that something is good, that it is valuable, and that there is a reason to pursue it—depends on the contents of one's cares. For example, whether Paul does *in fact* have a reason to cultivate his operatic talent, one might claim, depends on whether he cares about being talented at opera. A person who does not care in any way about being talented at opera does not have a reason to promote her operatic talent. This second thesis is about the nature of certain evaluative facts (in particular, it is a version of the Humean theory of normative reasons), and it is distinct from my main claim, which is about the nature of caring. Theorists who defend the centrality of caring for understanding the deep self are likely to endorse both claims, i.e., the thesis that caring entails dispositions to form certain judgments about the evaluative facts as well as the thesis that these facts themselves depend on the contents of one's cares. I wish to keep these claims separate, however. It is the former claim that is relevant, as my aim here is to characterize the nature of cares.

[14] These ideas are insightfully developed in David Shoemaker's "Caring, identification, and agency" (2003).

disapprobation, and despair. In this respect too, cares are quite different than desires. It is perfectly possible to desire something, but not have this rich and distinctive profile of emotional connections to the prospect of that thing being threatened, achieved, or foreclosed.

So cares involve a complex syndrome of motivational, commitmental, evaluative, and affective dispositions. What, then, is the connection between a mental state's exhibiting this syndrome and one's deep self? In particular, could it not be the case that a psychological state exhibits the full syndrome associated with cares and yet the attitude fails to be part of one's deep self?

In answering this question, notice that there is a fundamental connection between the functional role associated with cares and the conditions under which, intuitively, something genuinely *matters* to a person. When something matters to a person—when it is of real importance to her—she desires this thing in a non-instrumental way; she wants it for its own sake. She is committed to the thing. She resolves to face down environmental and psychic obstacles that get in her way and would experience a great sense of loss if forced to cease pursuing the thing. She judges having or attaining the thing to be good and valuable. Her emotions are bound tightly to the fortunes of the thing. It is simply not possible that these descriptions are all true of her and yet the thing in question is unimportant to her, or worse, that she is alienated from the thing. These observations suggest that there is a basic *conceptual tie* between the syndrome of dispositional effects associated with cares and *what it is* for something to matter to a person.

Now, a deep self contains the attitudes that make up one's fundamental conative point of view, the attitudes that specify what is genuinely important to the person. If the syndrome exhibited by cares has a basic conceptual tie to what it is for something to matter to a person, then this makes a strong case for the conclusion that one's deep self is constituted by one's cares.[15]

2.2.1 *Ontological versus psychological questions*

The question addressed by the care-based conception of the deep self is an *ontological* one: Which are the states that are *in fact* elements of one's deep self? The answer provided by the account is that the deep self consists of the set of one's

[15] This way of linking the deep self to one's cares uses a *direct* strategy. There is a direct conceptual tie, it is claimed, between the syndrome of dispositions associated with cares and one's fundamental self. Another plausible way to link cares with the deep self relies on an *indirect* strategy: Cares, perhaps only contingently, support some property *X*, where *X* is not part of the syndrome that characterizes cares. It is then argued that any state that plays this role with respect to *X* is part of one's deep self. One version of the indirect strategy comes from Agnieszka Jaworska, who in turns derives key premises from Michael Bratman. Jaworska argues that cares help sustain certain forms of cross-temporal continuities and connections, which on a broadly Lockean theory of personal identity are the basis of the agent's enduring identity over time. She then invokes Bratman's influential claim (Bratman 2000) that any state that plays this role in sustaining cross-temporal continuities and connections must necessarily belong to the person in precisely the way that is characteristic of elements of the deep self. See Jaworska (2007) for a detailed exposition of the steps of this argument. The direct and indirect strategies are interlinked and ultimately complementary. I believe both will figure into a comprehensive defense of the claim that one's deep self is constituted by one's cares.

cares. One's cares, like any other mental state, are in turn characterized in terms of the functional role properties they exhibit. So there is always a fact of the matter about whether a mental state is a care: If a mental state possesses the syndrome of dispositional properties that define what it is to be a care, then it is a care. Otherwise it is not.

There is a related *psychological* question of what states a person *takes* to be deep; that is, what states does a person regard as fundamental to her self?[16] The two questions are related because, more often than not, people are not self-deceived; the attitudes they take in this psychological sense to be deep are in fact deep in an ontological sense. But it bears emphasis that, on the care-based view of the deep self, a person can always be mistaken about the contents of her deep self. What makes a mental state a deep attitude is its exhibiting the relevant suite of functional role properties. It does not matter what the person thinks about the state, whether she regards it favorably, whether she consciously identifies with the state, or whether she recognizes the attitude as deep.

Recall Watson's man given a puritanical upbringing (see Sect. 2.1). He may take himself to genuinely care about sexual propriety and to reject certain kinds of physical gratification. But he is simply wrong. There is no state in his mental economy directed at sexual propriety with the characteristic functional role properties of a care. Now, it might well turn out that there *is* a care of the man's directed at pleasing his family and church community, which explains his having puritanical reactions. The man took his having these reactions as evidence that he genuinely cares about sexual propriety, but he was simply mistaken.[17]

A deep self theory of moral responsibility says that you are morally responsible for an action only if it expresses your deep self. The preceding distinction between ontological and psychological questions clears up what might be seen as an ambiguity in this formulation. It is the ontological sense of one's deep self that matters for the theory. You are morally responsible for your actions that reflect the person you really are, the actual content of your self. The states you psychologically identify with—that is the states you *take* to be fundamental to your self—are not, at least in any direct way, the basis for moral responsibility.

[16] This distinction is based on Agnieszka Jaworska's related discussion of ontological versus psychological senses of internality. See Jaworska (2007).

[17] Importantly, for a mental state to be a care, it must possess the syndrome of dispositions discussed in Sect. 3.1. It is consistent with a person's possessing these dispositions, however, that they are not actually *manifested* in a person's psychology—for example, a person has not *actually* exhibited commitment to the care, made care-concordant evaluative judgments, or experienced emotions tied to the fortunes of the cared-for thing. This reflects a general truth about objects defined by their dispositional properties. For example, a poison is something that, when ingested, disposes the person to die. If Smith ingests some compound *P* that is a poison, but subsequently takes an antidote so he does not die, *P* does not thereby cease to be poison. Objects retain their dispositional properties in the presence of antidotes, masks, finks, and other kinds of "defeaters" that prevent the dispositions from being manifested (see Cross 2011). So too in the case of cares; a person can care for some thing even if the syndrome of dispositions that defines what it is to care for that thing are, owing to the presence of defeaters, never actually manifested.

3 When does an action express one's deep self?

In this part of the paper, I put forward an account of the expression relation, i.e., the relation that must hold between a person's deep self and her action such that she is morally responsible for that action.

3.1 Agentially demanding views of expression

I start with an approach to how the notion of expression works that I think is on the wrong track, but studying why it falters is nonetheless instructive. This approach to the expression relation says that certain highly demanding agential conditions must be met in order for expression of the self to occur. Consider the endorsement-based account of expression found in the early writings of Frankfurt.[18] In the Frankfurtian "hierarchical" framework, agents have first-order desires, i.e., desires to do this or that, as well as various higher-order desires directed at other desires located in the motivational hierarchy. Endorsement involves the presence of the appropriate mesh between a certain first-order desire and one's desires of higher-order. This conception of expression is agentially demanding because the formation of higher-order desires is itself conceived of as being a highly reflective enterprise. In particular, the agent adopts a standpoint in which she steps back from her existing motives, looks across the full range of her desires and reflectively criticizes them, and on this basis forms new higher-order desires about which of her existing desires should or should not be effective in action.

Another conception of expression that is highly agentially demanding is offered by Susan Wolf. She proposes that an action is attributable to a person's deep self if "…she is at liberty (or able) to both govern her action based on her will, and govern her will on the basis of her valuational system" (Wolf 1993, p. 33). Wolf follows Gary Watson in understanding a valuational system as the rational faculty that issues in judgments of the form, "All things considered, X is the thing to do." While Wolf doesn't give an account of what it is for the valuational system to *govern* one's will, this certainly seems to imply a relatively strong form of counterfactual control—the person's will reliably complies with whatever her valuational system dictates.

Agentially demanding conceptions of expression such as these encounter serious problems. One problem, articulated forcefully by T.M. Scanlon and Angela Smith,[19] is that we routinely take people to be morally responsible for various kinds of spontaneous, non-deliberative conduct. Examples they cite include forgetting birthdays and anniversaries, attending to things one shouldn't, spontaneously finding certain (perhaps inappropriate) things amusing, and spontaneously experiencing emotions such as contempt, jealousy, or indignation. Notice this conduct does not appear to be reflectively endorsed or governed by one's valuational system.

[18] Frankfurt provided a number of different accounts of the conditions under which an agent endorses, or is identified with, a motive. I am focusing here on Frankfurt's earliest model, most clearly articulated in *Freedom of the Will and the Concept of the Person* (Frankfurt 1971).

[19] See Scanlon (1998), Smith (2005, 2008).

It certainly was not endorsed or governed in this way in the *present* as the relevant attitudes are *spontaneous*. Also, the conduct doesn't appear to have been endorsed or valuationally governed in the *past* either.[20] Common sense says we can be, and indeed often are, morally responsible for an enormous variety of spontaneous conduct. Indeed, it seems most natural to say that we are responsible for this conduct *precisely because* such conduct is deeply expressive of our selves. But agentially demanding approaches to the expression relation have trouble explaining this.

Another family of problems for agentially demanding views has been discussed in a series of works by Nomy Arpaly and Timothy Schroeder. I focus here on Arpaly and Schroeder's discussion of the case of Huckleberry Finn.[21] On their interpretation of the story, Huck reflectively judges that all things considered, he should turn in his friend, the runaway slave Jim. In the end, however, Huck just can't bring himself to do this. When a pair of slave hunters inquire about Jim, even as Huck continues to judge that he should turn Jim in, Huck spontaneously tells a lie to get the slave hunters to leave. If the preceding agentially demanding views of expression were correct, then Huck's action would not be expressive of his self and he could not be morally responsible for it. But this seems incorrect. We naturally think that Huck's action is profoundly self-expressive; it reflects Huck's caring for Jim and his respect for Jim's humanity. Moreover, there is a strong intuition that Huck is very much praiseworthy for what he does, something that would be impossible if he were not morally responsible for what he does. If this is right, then meeting agentially demanding conditions can't after all be necessary for self-expression; there must be routes to expressing one's self that bypass these conditions.[22]

Agentially demanding approaches to self-expression additionally face a third family of problems: They have difficulty accounting for certain "agents at the margins",[23] for example young children and people with certain disabilities, who appear to be able to express their selves in action. Consider a young child's

[20] Some philosophers defend *tracing conditions* for moral responsibility: A person is morally responsible for her spontaneous non-deliberative conduct because at some prior time she chose to develop or maintain the psychological mechanisms that are the basis for this conduct. Tracing has been persuasively criticized by Vargas (2005) [but see Fischer and Tognazzini (2009) for a response]. In my view, in addition to Vargas' critiques, the problem for tracing approaches is not *whether they can extensionally* capture our pattern of responsibility judgments; tracing conditions tend to be sufficiently vague and flexible that they usually can. The problem is that the *bases* for responsibility judgments that they specify seem incorrect (again, even if extensionally adequate). That is, when we assess whether someone is morally responsible for spontaneous conduct—for example, spontaneously and unthinkingly using a racial slur—the conditions in the remote past adverted to in tracing accounts seem to play no role in these assessments. What matters for moral responsibility is that the person's conduct is expressive of her self *in the here and now*. Relatedly, some may argue that to be responsible for the conduct flowing from one's self, the person must be responsible for shaping the contents of her self. Elsewhere, I raise doubts about whether responsibility for one's self is in fact possible. See Sripada (under review, sec. 2.7).

[21] See Arpaly and Schroeder (1999). The first chapter of Arpaly (2003) presents a number of additional interesting cases that vividly demonstrate a gap, and sometimes outright opposition, between what a person reflectively endorses and what is an expression of that person's self.

[22] I discuss the Huckleberry Finn case in more detail elsewhere, see Sripada (in press).

[23] See Shoemaker (2009) and Jaworska (1999, 2007) for lucid and poignant discussions of this topic.

comforting a crying parent; a partially senile woman's enjoying helping to prepare a meal for her family; an autistic man's spending the whole month continuously reading about dinosaurs.[24] These agents appear to have things that they genuinely care about, and their respective actions seem to be expressive of their cares. Since these agents lack sophisticated reflective capacities or the strong abilities for valuational governance that agentially demanding approaches require, these approaches have trouble explaining why these agents' actions are self-expressive.

3.2 The motivational support account of expression

What we need is a way to capture the idea that a person's self "stands for" or "stands against" an action[25] in a way that does not appeal to agentially demanding features such as endorsement, deliberate choice, or strong forms of valuational governance. In formulating an alternative "leaner" conception of the expression relation, consider a case. Raymond is a graduate student and moves into a seven-bedroom house with students from other departments. He immediately finds himself attracted to one of his new housemates, Millicent. A few weeks later, he is deeply in love with her. Raymond now cares for Millicent in a way that he didn't before he moved into the house and fell in love, and his newly emergent care exerts wide-ranging effects on his psychology. His actions change—he finds himself hanging around Millicent all the time and doing things to make her smile. His evaluations of prospects change. In weighing the options of going to the movies with Millicent or relaxing at home, the first prospect immediately strikes him as more attractive. His patterns of spontaneous conduct also change. He notices even the smallest signs that relate to Millicent, attends to topics that interest her, feels suddenly happy when she happens to show up, misses her when she is away, and feels bad when she is sad or disappointed. Raymond's caring for Millicent is implicated in all these changes; the presence of this new care influences his other psychological attitudes, comprehensively affecting his actions, attention, and emotions. Moreover, this newly emergent pattern of actions and spontaneous conduct is also, intuitively, deeply expressive of this care. This example suggests that the pattern of motivational influences exerted by a care on one's wider network of psychological attitudes has something important to do with expression. I take up this idea in proposing a *motivational support* account of expression in what follows.

In discussing the example of Raymond, I appealed to the idea that cares exert *causal influences* on one's various other attitudes and on action. It is tempting to infer that expression of some care *C* in one's action *A* simply requires that *A* is *caused by C*. Further reflection, however, suggests that this account is too simple.

[24] The first two examples are loosely drawn from Jaworska (2007).

[25] I am interested in formulating an account of moral responsibility that captures responsibility for both actions in the traditional sense (i.e., intentional actions) as well as various forms of spontaneous conduct such as noticings, attendings, and emotings. To aid exposition, going forward I use "action" throughout in an inclusive sense that encompasses both intentional actions as well as these various forms of spontaneous conduct. For emphasis and clarity, I sometimes still occasionally say "actions and spontaneous conduct", even though according to my usage, the former subsumes the latter.

Suppose Jimmy's son has gone missing in Afghanistan. He cares for his son so much that he ruminates continuously, and this in turn gives him a severe headache for which he must take an aspirin. Standard theories of causation would say that Jimmy's caring for his son causes his taking an aspirin—very roughly there is a chain of causal dependence that links the two. Jimmy's taking the aspirin, however, does not express his caring for his son. It seems then that the mere presence of a causal relation between a care and an action is not enough and a more refined account of the way that the relevant action is caused by the care is needed. Let me sketch such an account.[26]

Our minds are densely populated with *action-directed psychological mechanisms*, i.e., mechanisms that issue in actions or various kinds of spontaneous conduct. This class of mechanisms is extraordinarily heterogeneous. For example, it includes certain mechanisms that are relatively slow, conscious, and deliberative as well as other mechanisms that are relatively fast, automatic, and intuitive. One's various motives (i.e., one's cares, goals, and desires, etc.) exert *motivational influences* on these mechanisms, shaping the ways they unfold and favoring certain actions over others. Building on this idea, my proposal is the following:

> **MS**: An action *A* expresses a motive *M* if and only if during the operation of the action-directed psychological mechanisms that are involved in the etiology of *A*, *M* exerts motivational influences (of sufficient strength) in favor of *A-ing*.

Given the care-based conception of the deep self I defended earlier, an account of *expressing one's self in action* naturally follows:

> **ESA**: An action expresses one's self if and only if the motive expressed in the action is one of one's cares.

MS requires further clarification because it appeals to the notion of *motivation*. Let me spend some time discussing how I understand this notion. In any action-directed psychological mechanism, we can distinguish mental states with belief-like functional roles and desire-like functional roles. In particular, making a crude but useful simplification, we can distinguish these states in terms of *direction of fit*.[27] For both belief-like states and desire-like states, a perceived discrepancy between their contents and the state of the world disposes certain changes aimed at reducing this discrepancy. For belief-like states, the relevant dispositions involve changes to the contents of the belief-like states themselves; these contents tend to be adjusted to more closely match the world. For desire-like states, in contrast, the relevant

[26] In an insightful article, Levy notes limitations in previous attempts to understand the expression relation (Levy 2011). He proposes an alternative that says expression involves an attitude's causing an action "in the right sort of way," adding "there is a nonaccidental and direct relationship between the content of the action and the agent's actual attitudes" (p. 248). In my view, Levy's proposal is suggestive but ultimately underdescribed. One wants to know more about the nature of this direct and non-accidental connection between actions and the attitudes they express. The Motivational Support account of the expression relation that I propose might very well be thought of as an answer to Levy's challenge to propose a more adequate account of expression than thus far has been on offer.

[27] See Smith (1987).

dispositions involve the initiation of actions; the states of the world tend to be adjusted to more closely match the contents of the desire-like states.

Motivation refers to the dispositions towards action that desire-like states produce.[28] There are two fundamental properties of motivation. The *strength* of motivation refers to the degree to which the state disposes action. The *direction* of motivation refers to whether the state causally favors, disfavors, or is neutral with respect to the performance of a certain action.

All of this is pitched at a fairly high level of generality. If we are to say anything more specific about motivation—how it arises and the specific psychological changes it produces—then we must provide more details about the structure of the action-directed psychological mechanism at issue. This is the topic to which I now turn. I will discuss four importantly different examples of action-directed psychological mechanisms. My discussion will serve the additional purpose of illustrating the diversity of ways that a motive, by providing motivational support for an action, can thereby be expressed in that action.

The first example concerns deliberation about how to fulfill one's motives. Consider an action-directed psychological mechanism with three components: representations of various candidate actions that can be undertaken, representations of outcomes that the actions are believed to bring about, and various motives that are potentially satisfied by performing the actions. The mechanism assigns evaluative weights to candidate actions by figuring out how the actions (or more specifically the outcomes they bring about) stand in relation to the motives. When the assignments are complete, the highest ranked action is selected for implementation.

In a mechanism of this sort, an individual motive can influence which action is issued by affecting the evaluative weights that are assigned. If the person has a motive directed at achieving X, then actions that further X are assigned greater evaluative weight. If the relevant motive opposes Y, then actions that have Y as a consequence lose evaluative weight. Suppose Juan deliberates about whether to stay at work or go to his son's Little League game, and he ends up doing the latter. During deliberation, Juan's caring deeply about his son exerts motivational influences in favor of his going to the game. That is, via the assignment of evaluative weights, this care inclines the reasoning process in favor of this action. Thus according to **MS**, Juan's going to the game expresses his caring for his son, and it follows from **ESA** that his action is self-expressive.

A perhaps much more common form of action-directed deliberation employs *emotions* and *affect* as ways of evaluating prospects. When we care about something, situations in which the care is satisfied are imbued with a positive affective quality, or as I shall put it, they are "affectively marked" in a positive way. The assignment of affective markers to situations is not something we intentionally decide to do. Rather, it is something that our affective systems automatically and continuously engage in without the need for conscious awareness or supervision.[29]

[28] See Mele (1998, 2003).

[29] See Railton (2014) for a lucid, philosophically-focused discussion of this point.

Consider 10-year-old Reza, who is told by his mother to turn off the television and go play outside. He walks to the field and happens to meet young Elian, who just moved in next door. They play with trucks, tell jokes, and climb trees. Reza feels a warm inner "glow" when playing with Elian, and his affective system is responsible for this. Like most of us, Reza cares for having an affiliative connection with others (even if he is too young to articulate what "caring" or an "affiliative connection" is), and given this care, Reza's affective system recognizes the current situation as one in which the care is being satisfied. The next day, Reza's mother asks him whether he wants to watch television or go outside. As Reza considers the options, the prospect of his playing with Elian spontaneously strikes him as attractive (even if he can't say exactly why it strikes him this way). This is once again because of the operation of Reza's affective system, this time operating on *prospective* situations rather than actual ones. He goes out and plays. In this example, one of Reza's cares, via the production of affective markers, exerts motivational influences on his deliberation processes in favor of playing with Elian. According to **MS**, his playing with Elian thus expresses this care, and it follows from **ESA** that his action is self-expressive.

The third example of an action-directed psychological mechanism concerns *reinforcement learning*. This process also relies on affective markers, not to influence prospective deliberation, but rather to guide the learning of habits and other forms of spontaneous conduct. Consider Yuko, who has just moved to the Midwest. One day she just happens to smile when she crosses paths with a stranger. The stranger's face brightens and he beams a smile back at Yuko. Because Yuko cares about others' happiness, she experiences a burst of positive affect. This in turn reinforces her tendency to smile whenever she encounters new people. In this example, Yuko's caring about others' happiness inclines the relevant reinforcement learning process in favor of learning this pattern of smiling behavior. Thus according to **MS**, her smiling expresses this care, and it follows from **ESA** that her smiling is self-expressive.

The fourth example of an action-directed psychological mechanism differs somewhat from the previous three, as the relevant "action" in this case is the occurrence of an emotion. We sometimes think of emotions as entirely passive occurrences. They simply happen to us unbidden, irrespective of what we want. This perspective is, however, deeply misleading. While we typically can't exert much *control* over the initiation of an emotion episode—that is, we can't turn an emotion on or off by fiat—we are still active with respect to our emotions in the sense that our own underlying motives play a decisive role in their occurrence. Let me expand on this point.

It is a point of fairly wide agreement among philosophers and psychologists that emotions are initiated by *appraisal processes*—fast, automatic inferential processes that assess the significance of ongoing events with respect to one's motives.[30] Importantly, emotion appraisals are *independent* of one's deliberately formed, conscious, "person-level" judgments—notice that a person can consciously judge

[30] See, for example, Solomon (2003) and Ellsworth and Scherer (2003).

that flying is perfectly safe but still emotionally appraise the situation as one involving the threat of great harm. This divergence is possible because appraisals are implemented by proprietary processes that are distinct from those that subserve one's person-level judgments. Distinct emotions arise from distinct appraisals. Examples of appraisals associated with specific emotions include the appraisal for fear (the situation involves the threat of physical or mental injury to the agent), moral anger (the situation involves a normative transgression directed at the agent or her intimates), and jealousy (the situation involves someone's having or taking something that should belong to the agent). As in the previous examples of action-directed psychological mechanisms, the presence of a specific individual motive can incline appraisal processes to unfold in certain directions rather than others.

To see how emotional appraisals operate more concretely, consider the following case:

> *Jealous Director*
> A selfish father cares only for his professional success as a film director. His son adulates him and chooses the same career. One evening, they attend an awards ceremony where the father fully expects to win several prizes. Unexpectedly, it is the son who wins a prestigious award. Rather than feel happy for the son, the father feels only profound jealousy.

As I see it, here is the sequence of events that leads to the father's having the episode of jealousy: (1) The father cares deeply (and selfishly) for his own professional success. He cares about outshining all those around him, including his own son; (2) This selfish care influences his (fast, automatic) emotional appraisal processes, inclining these processes to "see" the current situation in terms of his son's having or taking something that should belong to him; (3) Once this appraisal of the situation is made, the other components of the jealousy syndrome (the cognitive, physiological, and phenomenological changes associated with jealousy) unfold. In this three-step sequence, the father's motivational attitudes, in particular his selfish caring for his own success, play a decisive role in the emotion appraisal process, strongly inclining these processes to produce an occurrence of jealousy. Thus, according to **MS**, the father's jealousy expresses this care, and it follows from **ESA** that his jealousy is an expression of his self.

These four examples of action-directed psychological mechanisms are certainly not exhaustive. They are nonetheless sufficient to illustrate my general point: Motives influence the operation of action-directed psychological mechanisms of various kinds, and when they provide the appropriate forms of motivational support for the doing of certain actions (or for the occurrence of certain forms of spontaneous conduct), then these motives are thereby expressed in these actions and conduct. In addition, these examples show that the Motivational Support account can make sense of how a *wide variety* of actions and conduct can be expressive of our selves, including actions that arise from explicit forms of deliberation, actions that arise from deliberation that involves affective markers, non-deliberative responses such as habits, and various kinds of spontaneous emotions. As I noted earlier, there is strong reason to believe that we can be morally responsible for all

these different kinds of action and conduct, and it counts in favor of the Motivational Support account that it readily explains this.

It is worth emphasizing a point that came up in the preceding discussion but should be made more explicit. On the Motivational Support account, expressing a care needn't involve having that very care *consciously* in mind when one acts. Recall that when Reza sees the prospect of playing with Elian as attractive, he isn't consciously aware of why he has this response—the relevant affective routines that guide this response operate below the level of awareness. A similar point applies to Yuko's habit of smiling. Indeed, not only might Yuko not know quite why she smiles at strangers, we can imagine a version of the case in which Yuko is not consciously aware that she is smiling at all (if shown a video of her social behavior, she might react with sincere surprise). So the Motivational Support account allows that motives can support actions through channels that are outside of awareness, and these actions can nonetheless be fully expressive of our selves.

The preceding point is important to keep in mind when considering a common objection to deep self theories of moral responsibility. These theories, it is claimed, can't make sense of responsibility for actions done "on a whim". Suppose Jill spontaneously decides to go to the amusement park. She may claim that the idea just popped into her head out of nowhere—for no particular reason it just seemed attractive to her. Her going to the park thus doesn't stem from any deeper motives and thus can't be an expression of her self. However, given the examples of Reza and Yuko, we should be suspicious of Jill's claim. Their respective affective systems operated outside conscious awareness to mark prospects as attractive. Similarly, it is likely that despite what Jill thinks, there is after all something about the prospect of going to the amusement park, perhaps camaraderie with friends, the chance to have new experiences, or the sheer joy of going on rides, that explains why it struck her as attractive. Jill's whim might thus not be so unmoored after all. It is likely to have its etiology in things—camaraderie, exploration, or joy—that she in fact cares about.

3.2.1 Priorities

We sometimes say that an action is expressive not of something a person *does* care about, but rather what he *fails* to care about, his attitudes of disregard or indifference. Consider a young, immature father who just doesn't care very much for his baby girl. He spends his time going out, partying, and socializing and neglects the interests of his child. Suppose that one evening, this man goes to a dance club event with his friends instead of attending his daughter's birthday party. Intuitively, his missing his daughter's birthday is expressive of his self. But how can the Motivational Support account make sense of this? It is hard to see how the man's disregard, i.e., the *absence* of a care, can provide motivational support for his action.

I propose that the way to understand these sorts of cases—i.e., cases involving disregard, indifference, and failures to care—is in terms of an agent's *priorities*. Notice that we naturally don't think of the young father as having two independent and unrelated attitudes: he cares about his partying lifestyle and, as a completely

separate matter, he disregards the interests of his daughter. Rather, it is more natural to see these two attitudes as intimately linked in the form of a settled ranking: in terms of how to spend his limited time, his partying lifestyle is far more important to him than being with his daughter.[31] Once we recognize that the man has a problematic *pair* of attitudes linked in the form of a prioritization, the problem raised earlier for the Motivational Support account goes away. It makes perfect sense that the man's problematic prioritization that places partying far ahead of his daughter provides motivational support for his going to the dance club and his missing his daughter's birthday. His missing his daughter's birthday thus expresses this problematic prioritization, and according to the Self-Expression account, he is morally responsible for what he does.

In some cases, such as the Neglectful Father case just discussed, an agent's prioritizations help to explain why his actions *are* self-expressive and why he is morally responsible for it. In other cases, an agent's prioritizations instead show why his actions are *not* self-expressive and why responsibility is mitigated or erased. As an illustration, consider the Unwilling Addict. When he is at last conquered by the desire for narcotics and uses the drug, he receives the pleasure of being high. If receiving pleasure is something that he, like most of us, cares about, then, on the Motivational Support account, isn't his action thus expressive of his self? I believe the answer is no. On the most plausible reading of this case, the Unwilling Addict has a prioritization in which certain ends that are incompatible with his drug-using lifestyle—his health, welfare, and meaningful relationships with others—are far more important to him than the kinds of temporary pleasures that he can get from the end of a needle; the former things matter more to him than the latter. Once we recognize the Unwilling Addict's prioritizations among drug-induced pleasures and other competing ends, then it is clear that his taking the drug does not express the attitudes of his self, and thus he is not morally responsible for it.

3.2.2 Failures of self-expression

Irresistible desires—such as those of the Unwilling Addict—provide one kind of case in which, on the Motivational Support account, expression of the self can fail to occur. There are others as well.

- Certain emotions can be non-self-expressive because they arise without a basis in one's cares. Consider an arachnophobe who sees her fear of ordinary house spiders as completely irrational. The things she cares about—for example, avoiding serious injury—don't play any role in supporting her spider-directed fears. Rather, her fear is directly triggered by the presentation of spider-ish

[31] What psychological facts determine the ranking of motivational attitudes within a prioritization? For example, in the case of the Partying Father, in virtue of what features of his psychology can it be truly said that the father prioritizes partying over the interests of his daughter? My very brief answer is that the relevant facts have to do with the functional role properties of the respective motivational attitudes that participate in the prioritization, that is, the syndrome of dispositions with which these attitudes are associated. Unfortunately, space does not permit a more detailed discussion of this important topic.

stimuli. In this case, the occurrence of spider-directed fear would not be expressive of her self.

- Severe forms of manipulation can inflict volitional damage and can prevent a person's self from motivationally influencing action. The heiress Patty Hearst, who was kidnapped and indoctrinated by a radical group, perhaps represents an example. Her armed robbery of a bank at the direction of her captors plausibly gained motivational support from the attitudes drilled in by these captors and not from her deep self.[32]

These examples show that the Motivational Support account allows for a number of different ways in which a person's self can fail to be expressed in action. It is notable that these are also cases in which, intuitively, the respective agents are not morally responsible for their actions.

To sum up, our goal was to formulate a leaner account of the expression relation that did not appeal to notions such as choice, endorsement, or strong forms of governance over action. The Motivational Support account of expression I have proposed does just this. Though it does not require meeting any highly agentially demanding conditions, it nonetheless captures when a person's self stands for, or stands against, what she does.

3.3 Comparing the motivational support account of expression with alternative views

To gain a deeper understanding of the more minimalist approach to expression relation that I have proposed, it is useful to contrast it with two alternative views: Frankfurtian endorsement views and so-called "could have done otherwise" views.

3.3.1 Frankfurtian endorsement

Frankfurt's notion of endorsement was one of the agentially demanding conceptions of the expression relation I briefly considered earlier, and it has been highly influential in the literature. It differs from expression on the Motivational Support account in a number of respects.

First, Frankfurtian endorsement is naturally construed as something an agent *does* (the agent steps back from her desires, reflectively criticizes them, and in some of Frankfurt's expositions, decisively commits to one among the competing desires). Expression on the Motivational Support account is not actional in this way. So long as one's cares play the right motivational role in supporting the doing of an action, then the action is expressive of those cares. There is nothing further the person

[32] Elsewhere, I provide a more detailed defense of the view that manipulation, when it undermines moral responsibility, does so by severing the expression of the manipulated agent's self in her actions (see Sripada 2012).

needs to do—no endorsements he must offer or decisive commitments he must undertake—in order for expression to occur.[33]

Another difference is that Frankfurtian endorsement demands sophisticated capacities for reflection and self-criticism, while the expression on the Motivational Support account does not. The latter only requires that the person cares for things, has the appropriate action-directed psychological mechanisms of the kinds discussed earlier, and these cares—during the operation of the relevant action-directed mechanisms—motivationally support the doing of the relevant actions. Arpaly and Schroeder's Huck Finn certainly meets these requirements and so the Motivational Support account explains why his action is self-expressive and why he is morally responsible and praiseworthy for what he does. Additionally, agents at the margin, who lack sophisticated reflective capacities, can certainly meet these requirements, at least in some contexts, explaining why in these contexts they can be morally responsible for their actions.

A third difference concerns the issues of consciousness and self-knowledge. Given the *actional* and *reflective* nature of Frankfurtian endorsement, it is hard to see how a person could fail to be aware of whether she endorses an action and how a person could ever be wrong about whether an action or attitude is in fact one she endorses. Neither of these is true on the Motivational Support account of expression. On this account, an action can express a person's self even if the person is not aware of this and even if the person herself is thoroughly convinced otherwise.[34]

The greater lenience of the Motivational Support account of expression greatly expands the range of conduct for which we can be morally responsible, in particular helping to explain why we can be morally responsible for spontaneous conduct, why Huck Finn-type agents are morally responsible for what they do, why young children and others with limited reflective faculties can be morally responsible for their actions, and why some implicit conduct that we are unaware of can nonetheless be an instance of moral responsible agency.

3.3.2 "Could have done otherwise" views

A popular and influential set of philosophical views say that to be morally responsible for an action, the person must have the ability to do otherwise. This requirement is usually cashed out in terms of a *counterfactual test*: A person is

[33] To be clear, what I mean is that expressing one's self is *itself* not an action (unlike endorsing one's motives), though of course one of the relata of the expression relation is an action.

[34] Matt King and Peter Carruthers have argued that Real Self views of moral responsibility (what I am here calling deep self views) are committed to the agent's meeting strong consciousness requirements. They further argue that current work in cognitive science suggests these consciousness requirements aren't likely to be met, thus perhaps undermining this family of views (see King and Carruthers 2012). Their critique applies, however, to versions of deep self views that have an agentially demanding *endorsement-based* approach to expression. The minimalist account of expression I have proposed as part of the Self-Expression account allows actions with a non-conscious etiology to be self-expressive. Thus these results from cognitive science, should they prove to be correct, shouldn't be taken to undermine *all* deep self views. I discuss relations between conscious awareness, moral responsibility, and deep self views in more detail elsewhere, see Sripada (in press).

morally responsible for *A*-ing if she would do otherwise than *A* were she to desire to, decide to, try to, etc.

These "could have done otherwise" views of moral responsibility have faced a host of serious counterexamples. One family of counterexamples concerns moral responsibility for emotions and other forms of spontaneous conduct. Consider the Jealous Director case discussed earlier. It is implausible that what makes the father morally responsible for his jealousy towards his son is that the occurrence of this emotion passes the relevant counterfactual test: The father would have "done" otherwise (i.e., he would not have felt jealous) were he to desire to, decide to, try to, etc.

Another family of counterexamples concerns agents whose morals and characters are such that it is impossible for them to do anything other than what they do. For example, Susan Wolf discusses a swimmer of impeccable character who sees a child drowning and without thinking any further jumps in the water to save the child. Given her thoroughly virtuous self, this person couldn't do otherwise than she does. Yet, intuitively, she is fully morally responsible for what she does.[35]

Finally, Harry Frankfurt presents the case of the Willing Addict. Recall that Frankfurt's Unwilling Addict has desires that are "too powerful for him to withstand, and invariably, in the end, they conquer him." The Willing Addict is similar in that his "addiction has the same physiological basis and same irresistible thrust" as the Unwilling Addict. However, he does not at all oppose his addiction:

> … [H]e is altogether delighted with his condition. He is a willing addict, who would not have things any other way. If the grip of his addiction should somehow weaken, he would do whatever he could to reinstate it; if his desire for the drug should begin to fade, he would take steps to renew its intensity (Frankfurt 1971, p. 19).

To make the case still clearer, consider a version in which the Willing Addict is completely unaware that he has an irresistible desire to use the drug. He loves using drugs so much that he has never tried to resist his drug-directed desires, and thus he has no inkling that he can't.[36] For many at least, the intuitive reaction to this case aligns with Frankfurt's view: the Willing Addict *is* morally responsible for what he does. After all, he does exactly what he deep down wants to do; why should it matter at all that, completely unbeknownst to him, he couldn't do anything else?[37] "Could have done otherwise" views, however, are forced to say that the Willing Addict is not morally responsible for what he does; in terms of moral responsibility, he is no different *whatsoever* than the Unwilling Addict.

The Self-Expression account of moral responsibility, which utilizes the Motivational Support account of expression, differs in a fundamental way from these "could have done otherwise" views: It does not require that we check various

[35] See Wolf (1993, pp. 58–59).

[36] See Frankfurt (1978, p. 160) for a description of a case along these lines.

[37] Elsewhere, I provide a comprehensive defense of the claim that the Willing Addict is morally responsible for his drug-directed actions. See Sripada (under review).

counterfactual scenarios to see whether the agent does something different than she actually does. Instead, what matters for moral responsibility is what happened in the actual scenario, and in particular whether the person's action received the appropriate kind of motivational support from her deep self. In the three preceding cases that presented problems for "could have done otherwise" views, i.e., the Jealous Director, Wolf's swimmer, and Frankfurt's Willing Addict, it is clear that the respective agent's selves did in fact—in the actual sequence that unfolded—provide the appropriate kinds of motivational support for his or her actions or conduct. This alone settles the issue of moral responsibility. In short, then, it is a distinctive feature of the Self-Expression account that all that matters for expression of the self, and thus for moral responsibility, is fully present and realized in the *actual sequence*.[38]

4 Refining the Self-Expression account

Having set out the Self-Expression account of moral responsibility, I now want to propose several refinements to how we understand the notion of a deep self and the notion of expression. These refinements better capture our ordinary commonsense understanding of the self and the ways that the self figures into morally responsible agency. They will also help to head off several objections that are commonly raised against deep self views.

4.1 Three refinements

I will introduce these refinements to the Self-Expression account through three pairs of distinctions. First, we can distinguish a *homogenous* conception of the deep self that denies that conflict within one's deep self can ever arise, versus a *mosaic* conception in which such conflict is allowed. On the homogenous conception,

[38] John Martin Fischer has proposed an influential "control-based" view of moral responsibility, see Fischer (1987) and Fischer and Ravizza (1998). On his view, the main criterion for moral responsibility, simplifying a little bit, says: A person is morally responsible for *A*-ing if, holding fixed the mechanism that actually issues in action, across a suitably broad range of counterfactual scenarios in which there is sufficient reason to do otherwise than *A*, the person would do otherwise than *A*. While Fischer's view and the traditional "could have done otherwise" views are in some respects different, there is also a key thing that they have in common: they both require an ability to do otherwise. The key difference is that on Fischer's view, the relevant ability is understood in terms of certain modal properties of the *mechanism that actually issues in action,* while on traditional views, the relevant modal properties belong to the *agent* [see Franklin (2014) for a lucid discussion of this point]. It is notable, and perhaps not surprising, that many of the same counterexamples to traditional "could have done otherwise" views appear to apply equally well to Fischer's view. In particular, it is implausible that the Jealous Director, Wolf's swimmer, or Frankfurt's Willing Addict have control in Fischer's sense over what they do, so Fischer's view has trouble explaining why they are morally responsible for what they do. Fischer sometimes calls his view an "actual sequence view" [see, for example Fischer and Ravizza (1998, p. 53)]. I am uncomfortable using this terminology to describe his view because his account employs a counterfactual test that is, in the end, highly similar to that employed in traditional "could have done otherwise" views. In contrast, views built along the lines of the Self-Expression account don't employ a counterfactual test, and hence it is more natural to refer to these as "actual sequence views".

conflict within one's deepest self always disappears on closer inspection. If two of a person's cares *appear* to conflict, then it must be the case that one of these is not *truly* a care; only one of the pair can ultimately be part of the person's self, while the other must be declared to be an imposter, external to the person in some way. I believe this view is incorrect, and it fails to appreciate the rich and sometimes painful complexities of being a human agent. On the mosaic conception that I favor, deep selves are potentially complex, heterogeneous things. They can encompass a variety of principals, commitments, and concerns that might be in subtle, or even not so subtle, tension with each other. To believe X, believe that Y is incompatible with X, and believe Y is irrational. To care for X, believe that Y is incompatible with X, and care for Y is not irrational. Rather, it is the human predicament; for us, conflict can and often does extend all the way to our very practical foundations. Sometimes a person has a single *prioritization* (see Sect. 3.2) among her competing cares. In other cases, the person is simply torn; two divergent cares, or even two divergent sets of prioritizations, pull the person, often tragically, in different directions, and each genuinely belongs to the person's self.

A second distinction that is sometimes intertwined but ultimately orthogonal is that between the *pure* versus *impure* conception of the deep self. The pure conception takes each person to care only about those things that are good or morally justifiable. I believe this conception of the deep self is mistaken. This conception confuses a person's deep self, the attitudes that specify what is *actually* important to the person, with her *ideal moral self*. The latter notion is, very roughly, the person who, when adopting a certain moral point of view, one would most want to be. The alternative impure conception of the deep self, which I believe is more reflective of commonsense understanding, allows that a person can genuinely care for ends that are morally dubious or even evil, or else genuinely care for ways of life that are unbecoming, unworthy of one's talents, or outright self-destructive.

Cartoonish villains might be impure in a thoroughgoing way. But in the real world, impurities in the self typically manifest as *flaws*. We pick out flaws in ordinary language with thick descriptors: vanity, self-indulgence, sloth, avarice, vindictiveness, and so on. While the attitudinal structure of flaws is complex, they invariably involve, as one of their constituent elements, impure cares and priorities, e.g., excessive self-love, pursuit of hateful ends, etc.[39] What further distinguishes flaws, which sets them apart from more thoroughgoing ways of embracing the bad, is that the cares and priorities that underlie our flaws represent only a small part of our self, and they are out of step with the remainder of our conative orientation. Put in more vivid terms, flaws are pockmarks on an otherwise smooth surface; they are small vortices within a larger, tranquil sea. Though they go against the ends that the remainder of our cares advance, flaws are not thereby external to our selves. On the mosaic conception that allows conflict within the self, and the impure conception that allows parts of the self to embrace the bad, our flaws remain fully part of who

[39] To be clear, I am not claiming that people care about their flaws, for example, that they care about being vindictive. Rather, the claim is that flaws such as vindictiveness have problematic cares and priorities as core elements.

we are. Our better angels and our demons oppose each other, and both belong to our selves.[40]

The third distinction concerns *how much of the deep self*, i.e., all/most of it or just some small part of it, must participate in the expression relation. Recall that on the mosaic conception of the self that I support, the self is made up of a diversity of cares that might be in mild tension or even open conflict with each other. On a *wide approach* to the expression relation, we try to figure out the overall stance of the deep self with respect to an action. This might be done by applying a formula that somehow "sums over" these various heterogeneous conative elements and arrives at a single overall conative orientation. An action expresses one's self if it stands in the right relation to this overall stance. On a *narrow approach*, we check the cares that constitute the person's deep self one by one. An action expresses one's self if it stands in the right relation to any one of these individual cares. Earlier, when presenting the Motivational Support account, and in particular in formulating the principle I dubbed **ESA**, I essentially assumed the narrow approach to expression is correct for a theory of moral responsibility. Let me now say a bit more about what justifies this assumption.

On the narrow approach to the expression relation, we judge a person as morally responsible when we find a suitable *anchor* in an individual element of the person's deep self. Now, finding an anchor of this sort for an action is compatible with a person's having *other* cares that strongly reject the action. For example, consider a married woman who is deeply in love with and devoted to her husband. But she has a flaw in her self, a subtle but insistent quality of self-love and vanity that leads her to seek out the attention and approval of other men. Suppose she finds herself flirting inappropriately with a handsome co-worker. On a wide approach to expression, her action does not express her self because it does not reflect the overall conative orientation that results from "summing over" all of the cares that make up her self. On a narrow approach to expression, her action does express her self because it finds a solid anchoring in the small cluster of attitudes within her self that constitute her flaw. The narrow approach to the expression relation thus finds support in our intuitions that, though she *also* cares for her husband and her marital commitments, the woman's action is nonetheless fully an expression of her self and she is morally responsible for it.

4.2 Out of character and weak-willed actions

I believe these three refinements are well supported by reflecting on our commonsense understanding of the nature of a person's fundamental self. In addition, the refinements give us the resources to address a family of objections that are commonly leveled against deep self views. Let me focus on two representative objections.

[40] The mosaic, impure conception of the deep self reflects, I believe, our folk psychological understanding of how the self is structured. To illustrate, here is historian Thomas DeFrank speaking about Richard Nixon: "He had demons. I mean I remember Jerry Ford once said to me Richard Nixon was 90 % a good person—there was 5–10 % of his persona that was bad and at times the bad just simply overwhelmed the vast majority of the good Nixon. And it usually came in situations … where he felt like he had been screwed his entire life by his political enemies and it was time for payback."

The first objection says that deep self accounts can't make sense of a person's being responsible for actions that are "out of character". To illustrate this objection, consider Jeeves, who is nearly always very deferential to his disagreeable boss. He does exactly as he is told without complaint, and indeed usually adds a servile "Yes sir, right away." This pattern of behavior expresses his self because he genuinely esteems figures of authority and cares about making them satisfied. But today, his boss insults him one too many times and in a particularly biting way. Jeeves angrily tells his boss to back off. Here Jeeves clearly acts out of character. The defender of the objection under consideration says that on a deep self view, Jeeves's action fails to reflect his deferential self, and thus he cannot be responsible for it.

The objection, however, is too quick. On the mosaic conception, selves can be complex, variegated things. Jeeves can care deeply about following along with authority and yet there can *also* be a part of him that is committed to maintaining a modicum of social respect, even if these two ends sometimes come into conflict. Furthermore, given the narrow notion of the expression relation, a person's action doesn't need to express all of a person's self; it can be anchored in small part—perhaps only a sliver. Given this, depending on the goal for which he acts, Jeeves's angry rebuke of his boss might very well be solidly anchored in a part of his self. Now, given Jeeves's track record of obsequious behavior, this would have to be a part that, as a matter of *statistical* fact, is not usually manifest in his actions, which is why we are correct in saying Jeeves acts out of character here. But the mosaic conception of the self coupled with the narrow approach to expression makes room for actions that are indeed out of character in this statistical sense, but still firmly tethered to one's self.

Notice that if Jeeves's reaction occurs in a totally blind rage, perhaps due to a mental disorder, and there is no anchoring for his action *whatsoever* in his self, then we should properly withdraw the claim that Jeeves is fully morally responsible for the action. We should instead say that Jeeves's action is not just out of character, but it is importantly alien to his self altogether, and thus moral responsibility is either mitigated or erased. This illustrates that deep self views not only can make sense of acting out of character, they can clarify when a person is morally responsible for doing so and when he is not.

A second commonly heard objection says that deep self accounts can't make sense of responsibility for weak-willed actions [see for example Fischer (2012, sec. 2.2) and Nelkin (2011, p. 17)], i.e., cases in which a person freely and intentionally acts contrary to her best judgment. The problem arises due to the following thesis:

(**W**) Weak-willed actions do not express one's deep self.

Since it is widely thought that a person is morally responsible for her weak-willed actions, then if **W** is true, it follows that deep self theories get the wrong results with cases of weakness of will.

Now, the defender of this weakness of will objection endorses **W** because she thinks if one's sincere all things considered judgment about the thing to do[41]

[41] Or some other kind of reflectively formed state, such as one's second-order volitions (see Fischer 2012).

opposes some action, this action cannot express one's self. But this is incorrect. To see why, notice that the mosaic, impure conception of the self allows that a person's self can be marred with flaws. For example, a dieter who cares deeply about her health and thus wants to stay on her diet may nonetheless have certain defects in her self: self-indulgence, laziness, gluttonous attachment to eating, prioritization of the present and a lack of care for her future, and so on. I believe that in cases of weakness of will, one's giving in to temptation is anchored in flaws in one's self such as these. In particular, either the temptation-directed action, the failure to regulate the desire that leads to this action, or both, are anchored in one's flaws, even if other parts of one's self are opposed, even very strongly, to giving in. Moreover, the presence of this sort of anchoring is precisely what distinguishes weakness of will from compulsion. For an individual who acts due to a true compulsion, the action is not anchored in her flaws, and indeed does not express any part of her self at all. In short, then, adopting a mosaic, impure conception of the self and adopting a narrow notion of expression help to show that **W** is false and that deep self approaches can after all accommodate moral responsibility for weak-willed actions. Indeed, adopting these refinements helps us to go further than just this. As in the case of out of character actions, these refinements help to illuminate the target phenomenon more fully, in this case by helping to draw an otherwise elusive boundary between weakness of will and compulsion.

My aim in this section has been to block certain commonly voiced objections to deep self views of moral responsibility. Thus, I have tried to show how the Self-Expression account, by adopting a mosaic and impure conception of the self and a narrow approach to expression, does in fact have the resources to deal with the kinds of actions discussed above that are often seen as problematic for deep self theories. Now, I haven't tried to *assemble* these resources into detailed, fully specified accounts. Thus, my discussion has left a number of fascinating and important questions not fully answered—for example, how to understand conflict within one's deepest self and how to think about flaws in one's self. A full account of these phenomena, however, is too much to be taken on here and so I leave these important topics for another day.

4.3 The cognitive approach to the deep self revisited

I believe that these three refinements, i.e., the mosaic and impure conceptions of the deep self and a narrow approach to the expression relation, are all highly plausible in their own right; they find strong support when we reflect on our commonsense understanding of the structure of the self and the role it plays in morally responsible agency. Moreover, these refinements were needed to explain moral responsibility for out of character and weak-willed action. I now want to examine these three refinements in light of the first "choice point" for a deep self theory that I identified earlier.

Earlier, I distinguished cognitive versus conative approaches to the deep self. I argued that cognitive views face a key *Which Judgments?* problem—among the myriad judgments we make across time and context, they need to demarcate which judgments count as part of the deep self and which do not. The preceding discussion

suggests an additional problem for cognitive approaches: These approaches don't appear to be easily combined with the three refinements just discussed.

The mosaic conception of the deep self says that the contents of one's self can be inconsistent and even in open conflict. A conative conception of the deep self readily accommodates this, as it is not just possible but indeed *typical* that one's set of cares contains elements that are starkly opposed. In contrast, it is not at all obvious that one's reflective judgments can be in open conflict in this way. A person who reflectively judges that *A*-ing is most worth doing and *at the same time* judges that *A*-ing is *not* most worth doing is not conflicted. Rather, if the idea can be made sense of at all, the person is simply irrational. The problem is actually more serious than this. Recall that in order to address the Which Judgments? problem, it appeared we needed to adopt the view that it is not just any old evaluative judgment that is part of the deep self. Rather, the deep self consists of only those evaluative judgments made in certain epistemically favorable circumstances. It is especially hard, however, to make sense of one's epistemically idealized judgments exhibiting synchronic divergence in the ways envisioned with the mosaic conception of the deep self.

The impure conception of the deep self too seems hard to combine with cognitive approaches to the deep self. One important reason is that, once we set cartoonish super villains aside, impurities in the self typically manifest as flaws—small islands of impurities surrounded by a larger pool of deep attitudes that are opposed in content. Flaws make sense only on the mosaic conception of the self that allows conflict in one's fundamental self, and cognitive deep selves, as we have just seen, sit poorly with mosaicism about the self. The narrow approach to the expression relation too depends on the mosaic conception of the self, as differences between narrow versus wide expression only emerge when there is some quantity of conflict in one's fundamental self.

In sum, if we want to hold onto these three refinements, i.e., the mosaic and impure conceptions of the deep self and a narrow approach to the expression relation, then it seems we are under strong pressure to reject cognitive approaches to the deep self. Now, it may be that there are clever ways in which these three refinements can be combined with cognitive deep selves. But until some suggestions along these lines are put forward, adopting these refinements (and as we have seen, there appear to be excellent reasons why they should in fact be adopted) must come at the cost of rejecting cognitive deep self views.

5 Conclusion

I have proposed a new deep self theory of moral responsibility, the Self-Expression account, that addresses problems with existing deep self views. What is a deep self? I have argued for a conative view in which one's self consists of one's cares. I gave an account of cares in terms of their functional role; they exhibit distinctive motivational, commitmental, evaluative, and affective dispositions. What does it mean to say an action expresses one's deep self? I proposed the Motivational Support account of the expression relation that—in contrast to alternative accounts

such as the endorsement approach—does not require meeting highly agentially demanding conditions. The Self-Expression account explains moral responsibility for a much broader array of conduct than existing deep self views. It also fits better with our commonsense understanding of the structure of the self and the role it plays in morally responsible agency.

Acknowledgments This work was supported by the John Templeton Foundation. Versions of this paper were presented at the Science of Ethics Workshop (Ann Arbor, MI, June 2012) and at the University of Maryland (College Park, MD, March 2013). Thanks to Sarah Buss, Peter Carruthers, John Martin Fischer, Alfred Mele, Edward Nahmias, David Shoemaker, Walter Sinnott-Armstrong, and Angela Smith for very helpful comments.

References

Arpaly, N. (2003). *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press.

Arpaly, N., & Schroeder, T. (1999). Praise, blame and the whole self. *Philosophical Studies, 93*(2), 161–188.

Brandt, R. B. (1998). *A theory of the good and the right*. Amherst, N.Y.: Prometheus Books.

Bratman, M. E. (2000). Reflection, planning, and temporally extended agency. *Philosophical Review, 109*(1), 35–61.

Buss, S. (2012). Autonomous action: Self-determination in the passive mode. *Ethics, 122*(4), 647–691.

Cross, T. (2011). Recent work: Dispositions. *Analysis*, p. anr144.

Dewey, J. (1957). *Outlines of a critical theory of ethics*. New York, NY: Hillary House.

Ellsworth, P.C., & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences*. Series in affective science (pp. 572–595). New York, NY: Oxford University Press.

Fischer, J. M. (1987). Responsiveness and moral responsibility. In F. Schoeman (Ed.), *Responsibility, character, and the emotions: New essays in moral psychology* (pp. 81–106). Cambridge: Cambridge University Press.

Fischer, J. M. (2012). Semicompatibilism and its rivals. *The Journal of Ethics, 16*(2), 117–143.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. New York: Cambridge University Press.

Fischer, J. M., & Tognazzini, N. A. (2009). The truth about tracing. *Noûs, 43*(3), 531–556.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68*(1), 5–20.

Frankfurt, H. G. (1978). The problem of action. *American Philosophical Quarterly, 15*(2), 157–162.

Frankfurt, H. (2006). *Taking ourselves seriously and getting it right*. Stanford: Stanford University Press.

Franklin, C. E. (2014). Everyone thinks that an ability to do otherwise is necessary for free will and moral responsibility. *Philosophical Studies, 172*(8), 2091–2107.

Jaworska, A. (1999). Respecting the margins of agency: Alzheimer's patients and the capacity to value. *Philosophy & Public Affairs, 28*(2), 105–138.

Jaworska, A. (2007). Caring and internality. *Philosophy and Phenomenological Research, 74*(3), 529–568.

King, M., & Carruthers, P. (2012). Moral responsibility and consciousness. *Journal of Moral Philosophy, 9*(2), 200–228.

Levy, N. (2011). Expressing who we are: Moral responsibility and awareness of our reasons for action. *Analytic Philosophy, 52*(4), 243–261.

Mele, A. R. (1998). Motivational strength. *Noûs, 32*(1), 23–36.

Mele, A. R. (2003). *Motivation and agency*. USA: Oxford University Press.

Meyer, S. S. (2011). *Aristotle on moral responsibility: Character and cause*. Oxford: Oxford University Press.

Nelkin, D. K. (2011). *Making sense of freedom and responsibility*. USA: Oxford University Press.

Railton, P. (1986a). Facts and values. *Philosophical Topics, 14*(2), 5–31.

Railton, P. (1986b). Moral realism. *The Philosophical Review, 95*(2), 163–207.

Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics, 124*(4), 813–859.
Rosati, C. S. (1995). Persons, perspectives, and full information accounts of the good. *Ethics, 105*(2), 296–325.
Scanlon, T. (1998). *What we owe each other*. Cambridge, MA: Harvard University Press.
Sher, G. (2009). *Who knew?: Responsibility without awareness*. Oxford: Oxford University Press.
Shoemaker, D. (2003). Caring, identification, and agency. *Ethics, 114*(1), 88–118.
Shoemaker, D. (2009). Responsibility and disability. *Metaphilosophy, 40*(3–4), 438–461.
Smith, M. (1987). The Humean theory of motivation. *Mind, 96*, 36–61.
Smith, A. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics, 115*, 236–271.
Smith, A. (2008). Control, responsibility, and moral assessment. *Philosophical Studies, 138*, 367–392.
Solomon, R. C. (2003). I. Emotions, thoughts and feelings: What is a "Cognitive Theory" of the emotions and does it neglect affectivity? *Royal Institute of Philosophy Supplements, 52*, 1–18.
Sripada, C. (2012). What makes a manipulated agent unfree? *Philosophy and Phenomenological Research, 85*(3), 563–593.
Sripada, C. (in press). Acting from the gut: Responsibility without awareness. *Journal of Consciousness Studies*, *22*(7–8).
Sripada, C. (under review). Frankfurt's unwilling and willing addicts.
Strawson, P. (1962). Freedom and resentment. *Proceedings of the British Academy, 48*, 187–211.
Vargas, M. (2005). The trouble with tracing. *Midwest Studies In Philosophy, 29*(1), 269–291.
Velleman, J. D. (1988). Brandt's definition of "Good". *The Philosophical Review, 97*(3), 353–371.
Watson, G. (1975). Free agency. *The Journal of Philosophy, 72*, 205–220.
Wolf, S. (1993). *Freedom within reason*. New York, NY: Oxford University Press.