A New View of Self-Constitution

Abstract: In this paper, we criticize what we dub the "pruning view" of self-constitution and defend an alternative derived from psychoanalysis. We discuss the common core shared by different version of the "pruning" view. That core combines two main ideas: the first is a proposal about how to draw the boundary between self-proper and non-self, i.e., how to identify those elements of the self that are good candidates for "pruning." The second is a suggestion about how to handle the unwanted parts in an attempt to achieve a coherent self. With regard to the first point, on the pruning view, the self-proper is to be identified with some "higher" agential component of a person's mental economy, paradigmatically, reason and rationality, or else reflection. With regard to the second part, an agent on the pruning view can achieve the goal of an integrated self by giving affirmation to the central aspects of his or her self while refusing to license the undesirable parts.

We argue that normative assessment comes too early on the pruning view, so early that it interferes with achieving deeper self-understanding and producing lasting change. On the proposal we advocate, "understanding first," undesired and undesirable parts of the self are often an integral part of the self, properly speaking, and must be recognized and engaged with as such. Only after deeper self-understanding has been achieved is the attempt to liberate oneself from unwanted elements likely to succeed.

1. Introduction

We are not fully unified creatures. Occasionally, we have competing desires. For example, a person may want to have breakfast and continue to sleep at the same time. Or, we might consciously want one thing and unconsciously want another, as when an aspiring actress who thinks she does not deserve to succeed ends up sabotaging her own prospects. We may have desires that conflict with our better judgment, for instance, a CIA operative who judges that protecting classified information is paramount may nonetheless have a desire to confide classified information to a family member. Again, we may have irrational motives and fears, as in the case of an otherwise reasonable man who starts avoiding an officemate due solely to having had a dream in which this officemate hurts him. Finally, we may act out of character, for example, a generally polite and courteous person may inexplicably fly into a rage.

There is a commonsense view that major disunity and serious fragmentation within a person's self are not good. Thus, seeking to integrate the different facets of our selves is desirable (though it

is also typically ceded that striving for *perfect* integration is likely a bad idea; *some* fragmentation is an ineradicable part of the human condition[1]). But how is unity to be successfully achieved? On one popular philosophical picture, the answer is: identify the undesirable elements and then subdue and extrude them through self-control. Do not allow unlicensed parts of your psychology to become your will. It is precisely this process of making yourself into a unified agent whose mental life exhibits harmony which is sometimes referred to as "self-constitution." For ease of reference, we will call this type of account "the pruning view" of self-constitution, since on this view, inner harmony and unity are achieved by "pruning" discordant elements. Historically, we find predecessors of this view in Plato, Kant, and Joseph Butler, and more recently, in Kantian philosophers such as Korsgaard and Shapiro, and in another version (departing from Kant) – in early Frankfurt.[2]

We note that while disunity and fragmentation are widely perceived as quite undesirable, the antidote to disunity, self-constitution, can be thought paradoxical: it may seem that the self doing the constituting must *already* be a well-constituted self. Otherwise, how can it do the trick? In the words of David Velleman, it is "as if the rabbit could go solo and pull himself out of the hat." [3] The idea seems much less paradoxical, however, if we move beyond the label and explain what proponents of self-constitution typically mean.

The pruning view has different versions, but for present purposes, we are interested in the features those versions share.[4] The common core combines two main ideas: the first is a proposal about how to draw the boundary between self proper and non-self, i.e., how to identify those elements of the self which are good candidates for "pruning." The second part is a suggestion about how to handle the unwanted parts in an attempt to achieve inner harmony.

With regard to the first point, on the pruning view, the self proper is to be identified with some "higher" agential component of a person's mental economy, paradigmatically, reason and rationality (Plato, Kant) or else reflection understood as either the process of normatively assessing

one's first-order attitudes in light of reasons (Joseph Butler, Korsgaard, Shapiro) or as an ability to form second-order desires (early Frankfurt).[5] With regard to the second point, an agent, on the pruning view, can achieve inner harmony by giving affirmation to the central aspects of his or her self while refusing to license the undesirable parts. Here is how Frankfurt describes the process of "pruning":

> In rejecting the desire…the person withdraws himself from it. He places the rejected desire outside the scope of his preferences, so that it is not a candidate for satisfaction at all. Although he may continue to experience the rejected desire as occurring…the person brings it about …that its occurrence is an external one. The desire is then no longer to be attributed strictly to him, even though it may well persist or recur as an element in his experience.[6]

Korsgaard would object to Frankfurt's suggestion that the second-order attitudes relevant to self-constitution are simply second-order desires[7] and will insist that only *reason* can be the sort of "legislator" necessary for self-constitution. More generally, on Korsgaard's view, there can be no inner harmony without morality.[8] But her account is structurally analogous to Frankfurt's. Thus, endorsing what she takes to be Plato's view of the harmony of the soul, she says:

> Appetite presents the proposal; reason decides whether to act on it or not, and the decision takes the form of a legislative act. This is clearly the Constitutional Model.[9]

So for proponents of the pruning view, an agent engaged in self-constitution is a bit like a sailor on a boat who places his weight on the good parts of the boat in order to cut out and repair the damaged parts.

It would be difficult to exaggerate the prominence of the pruning view of agential self-constitution in philosophy – likely, every professor who has taught introductory ethics has shared some version of it with students, though perhaps, not everyone has used the label "self-constitution." Popularity here is hardly a historical accident: rather, it is testimony to the deep intuitive appeal of the

picture. There is something compelling about the old Platonic idea that the different fragments of our selves are not all on a par, that some – paradigmatically, reason – are more central to who we are than other parts, and that those central parts ought to be put in charge since they are identified with the person in a way in which the peripheral and "external" parts are not. Thus, we say of a person who cannot control his desire to gamble and ends up losing all his possessions that he is "controlled by an impulse," alleging that the unlicensed impulse to gamble, though it occurs *in* the agent, is somehow not to be *identified* with the agent and assails the agent, as it were, from without. The agent, we think, is the one who struggles to overcome the impulse while the impulse is a kind of alien force. If we thought, by contrast, that the impulse to gamble has a stronger claim on being identified with its possessor than reflective control does, we would, instead of saying that the struggling addict is controlled by a compulsive desire, say rather that the continent person – the agent who silences his desire to gamble – is "controlled by reason." But we exhibit a clear preference for the former way of speaking over the latter.

Of course, not all aspects of the self which are in tension with each other are "ordered" in the sense implied by the pruning view. That is, not all inner conflicts are conflicts between "high" and "low." Going back to the earlier examples, the competing desires to have breakfast and to continue sleeping at the same time are, on the face of it, equally central to a person's identity.[10] The irrational fear of an officemate, however, or the anger which comes unbidden and makes a person furious for no apparent reason are, arguably, lower down the hierarchy compared with reasoned judgments such as the judgment that rage would be an extreme overreaction in a given case or that a bad dream involving a colleague's being hostile toward us does not provide good grounds for shunning a friendly colleague.

Despite its intuitive appeal, the pruning view of self-constitution is not a consensus view among moral philosophers. In particular, the idea of the self which follows from the pruning view is

widely contested. Critics have argued that proponents of the view accord too great a role to reason and second-order attitudes and too little to conative elements and first-order attitudes. The thought is that pruners overemphasize the conscious, active role we play in making ourselves who we are and fail to recognize the importance of seemingly passive or even unconscious aspects of the self – those desires, cares, and character traits, weak-will, lapses of attention, etc., neither consciously chosen nor endorsed – in action and identity.[11] More recently, a family of alternatives aimed in part at remedying alleged problems with this type of account by placing an emphasis on unchosen, broadly conativist, aspects have been developed.[12]

There is much to be said for "conativist" views and their inclusivity. The self does appear to contain unchosen, unendorsed, and sometimes – unendorseable – aspects which are nonetheless an ineliminable part of who we are. But if what one is interested in is self-constitution, such views do not have much to offer. The very label "self-constitution" has been, not without justification, almost exclusively appropriated by proponents of the pruning view. What we would like to do in this paper is offer an alternative. The proposal we will advocate is derived from psychoanalysis, and it combines the inclusivity of (the most inclusive among) conativist approaches with the idea of self-constitution. We will defend this alternative on both theoretical and practical grounds. Briefly, we will argue that the formation of second-order volitions regarding one's first-order motives comes much too early on the pruning view, so early that it interferes with achieving deeper self-understanding and producing lasting change. On the proposal we advocate, undesired and undesirable parts of the self are often an integral part of the self, properly speaking, and must be recognized and engaged with as such. Only after deeper self-understanding has been achieved is the attempt to liberate oneself from unwanted elements – minimally to prevent acting upon them – likely to succeed.

In addressing our task, we proceed as follows. We begin with a brief discussion of the advantages and disadvantages of conativist views as compared to the pruning view (Section 2). Then

we introduce the approach we favor in the context of an example (Section 3). We proceed to ask whether our critique of the pruning account does not derive its force from an implausible view of externality. This leads us to consider the question of whether there are any aspects of the self that could be deemed "external." We suggest that the question of externality (to the self) has no simple answer and that the best way to address it is to *begin* with self-constitution: while others have sought to define the boundaries of the self and then, propose a view of self-constitution, we, by contrast, suggest that it is preferable to begin with a viable account of self-constitution and construe the self as whatever emerges in the process of self-constitution (Section 4). Having made our case for this way of proceeding, in the final section (Section 5) we present our conclusions and suggest that only after agential self-extending work can people constitute themselves as the persons they wish to be.

2. The conativist criticisms of the pruning view

There are two tacks to the criticism conativists, generally speaking, levy against their pruner opponents in an attempt to show that the pruners' account is too restrictive. The first line of criticism has to do with responsibility for action. Conativists allege that we may be responsible, praiseworthy and blameworthy for actions that issue from peripheral aspects of the self, including aspects with regard to which we feel completely passive, or which conflict with our better judgment. Thus, if someone is mean to another owing to an unchosen trait of meanness, or worse forgets a baby in a hot car through an involuntary lapse of attention, both the mean act and that of forgetting of the baby are attributable to the agents who perform them, and these agents are responsible for those actions. Similar considerations apply to actions for which we are praiseworthy: if a person heroically but unthinkingly risks her life to save another, she is responsible – and praiseworthy – for that act too, despite the act's automaticity and lack of deliberation[13] (and it may be that *most* heroic behavior is automatic in just that way[14]).

Proponents of the pruning view such as Korsgaard have responded to this concern by saying that bad actions which issue from (so called) passive aspects of the self are *defective* actions, i.e., not *real actions*. Korsgaard conceptualizes such actions as a kind of failure – a failure of self-constitution – and suggests that there is nothing puzzling about the thought that people may be responsible and blameworthy for their failures. Korsgaard and like-minded philosophers can argue further, in a Kantian vein, that when it comes to actions that appear heroic but are performed automatically and unthinkingly, those may be admirable in *some* sense, but they do not possess true moral worth and are akin to the good actions of lower-level animals.[15]

Even if the first set of criticisms can be blocked in this way[16], there is a second problem. It concerns not responsibility for action, but instead identity: *who* a person is. Suppose we agreed with proponents of the pruning view that unlicensed aspects of the self can be properly external to the core self yet still apt to give rise to actions – both good and bad – for which people are responsible. What remains unclear is this: In what sense can we be said to be capable of renouncing parts of ourselves, as per the pruning view? Terence Penelhum, for instance, giving voice to just this type of concern, says that the idea that one part of the self can disown another "is merely a metaphorical truth or a simple literal falsehood."[17]

The problem, as we see it, is that the self is complicated, and some of the desires we may wish to master, subdue, and prune may be deeply expressive of a person and therefore not legitimately subject to pruning, given the essentials of pruning.[18] Consider, for instance, the character of Margaret Gray from Rebecca West's novel *The Return of the Soldier*.[19] Margaret was once in love and engaged to be married to an attractive, upper-class man named Chris Bauldry, whom she has not seen in years. Meanwhile, Chris has gotten married, then went to the front. Chris comes back from the war with shell-shock amnesia. He has absolutely no recollection of anything that's happened to him in the past 10-15 years – no memory of having married his wife Kitty, a beautiful woman of exquisite taste. No

memory of his deceased son, Oliver. Chris remembers, instead, being in love with Margaret Gray, a woman he'd dated years earlier. He claims he cannot live without Margaret. His old love is now married to an unsuccessful innkeeper, but her old feelings for Chris are back with a vengeance. She begins visiting Chris Bauldry's estate every afternoon and spending time with Chris. Chris and Margaret appear to be blissfully happy together. A doctor named Anderson proposes to help recover Chris's lost memories through psychoanalysis. Margaret feels conflicted: at some level, she acknowledges that helping Dr. Anderson enable Chris to recover his memories would be the right thing to do: Chris would go back to his wife, and she to her husband. However, she loves Chris and wants to keep him to herself. But she must subdue this selfish desire in order to do the right thing. The desire to remain with her beloved Chris, however, is deeply expressive of who she is. If Margret decides to silence her desire to be with Chris, she may experience that as a painful sacrifice, as "killing" a part of herself. Returning to our boat metaphor, the experience will resemble not repairing damaged parts but rather, throwing valuable, even essential to long-term functioning objects, overboard to lighten cargo load.

The upshot here is that conativist arguments for inclusivity succeed. But what about self-constitution? How is a person on this type of picture supposed to *integrate* the different aspects of her psyche and "constitute" herself? If the conativist's answer is to offer the old "self-control" remedy, then, whatever the conativist's view of the self and of responsibility, her notion of self-constitution is more or less the same as those of proponents of the pruning view. Moreover, if the conativist is right in being as inclusive about the self as she is, it is not clear that she *could* give the "self-control" answer to the question. She cannot, because when the cognitive part of the self is in tension with those conative parts which are central to identity, self-control itself can be in tension with the goal of self-constitution. Take the case of Huck Finn. Arguably, if Huck had had greater self-control, he would have given up Jim. But he would not have, thereby, succeeded in the goal of self-constitution. He is

closer to success in acting just as he acts. Yet, if self-control is not the (simple) answer, what *other* answer is there?

One could, perhaps, say here, on behalf of the proponents of conativist accounts, that inner harmony is simply a matter of luck rather than an achievement of sorts. For of course, it is always possible that a constitutionally lucky person may *happen* to have psychic unity. But what about the not-so-lucky? The conativist could give a "partners in crime" reply and say that moral luck is a problem for all accounts, including those that construe inner harmony as an achievement of the will, since the strength of one's will is itself (at least partly) a matter of luck. But this reply is not satisfactory: First, intuitively, self-control does minimize the role of moral luck (although it does not eradicate it), so the moral luck problem is more pressing on the conativist than it is on the pruning view. Second, we should not resign ourselves to a (defeatist) moral luck line unless there simply isn't a better alternative. But there *is* a better alternative, or so we will argue.

3. Understanding first

We wish to introduce the main idea by means of an example, which we will discuss in some detail. We call it *Bully*.

Bully

Jacob has the trait of chronic bullying – he frequently tries to bully those around him, particularly those who are weaker. This has worked for him in high school where he was popular in some circles precisely because he was a bully, but now that he is in his late 20s, he finds himself increasingly ostracized and unpopular. The shunning of others makes him even more prone to bully, creating a vicious cycle. Jacob has come to realize that his behavior is maladaptive, but he just can't shake off the old habits.

On what we have called the "pruning view" of self-constitution, what Jacob must do, along with recognizing his desire to bully as bad, is to master and subdue his desire to bully others. There

are a few problems with this proposal. First, how is Jacob to recognize the badness of his desire to bully others? The standard answer would be: by means of critical reflection.[20] But reflection may serve to distort reality rather than help achieve clarity. Consider again the case of Margaret Gray. Margaret believes that the reason she opposes the idea of helping Chris recover his memories is *Chris's* happiness:

> "Doctor," she said, her mild voice roughened. "What's the use of talking? You can't cure him." She caught her lower lip with her teeth and fought back from the brink of tears, -- "Make him happy, I mean. All you can do is make him ordinary."[21]

The reader here can't escape the impression that it is her *own* happiness Margaret does not want to give up, not Chris's. Though she is probably right that amnesiac Chris is "less ordinary" than Chris without amnesia, she has no good reason to believe that Chris is happier now than he would be if he remembered his wife.

Real-world examples of the flawed nature of reflection are not hard to come by. In *Virtuous Violence*, anthropologists Alan Fiske and Tage Rai present compelling evidence that almost all violent criminals – people guilty of spousal and child abuse, rape, and murder – believe they are acting with good moral justification.[22] The authors have this to say to UCLA Newsroom's Meg Sullivan on the topic: "Except for a few psychopaths, hardly anybody harming anybody else is doing something that they intend to be evil … On the contrary, they intend to be doing something right and good."[23] The violent criminals studied by Fiske and Rai have deployed reflection with the result that they find themselves justified in engaging in horrendous acts.

Similar considerations apply to Jacob the Bully: Jacob may, when he comes to judge his bullying as morally wrong, seek to give morally acceptable descriptions to his actions, for instance interpret his bullying as a reasonable response to external provocation. Even patently irrational motives and fears could be endorsed and seen as being in accord with our better judgment, for

instance, a patient with OCPD really does believe that her elaborate rituals are necessary to prevent the world from collapsing.[24]

Richard Moran, responding to a point similar to ours made by Jonathan Lear (who, as we do, writes from a psychoanalytic perspective), offers a "partners in crime" rejoinder. Moran argues that, while reflective judgment can be misused, so can any other procedure, including therapeutic procedures. Lear contends that reflection may serve to preserve anger by convincing the angry person that she is not angry. [25] Moran counters:

> I am not sure just how we are to understand Lear's specific point that reflective judgment may serve to protect and preserve the anger or fear in question, for it seems to me that this is a general liability for any stance or procedure one may adopt with respect to some part of oneself, and is not restricted to a particular philosophy or therapeutic practice. Even in those traditions of philosophy that are explicitly therapeutic, such as the practices of Socrates, or Nietzsche, or Wittgenstein, such a risk is endemic to their procedures. Socratic elenchus can itself be pressed into the service of preserving ignorance, Nietzschean genealogy can be practiced in the spirit of the ascetic ideal, and the practice of Wittgensteinian therapeutic reminders can be distorted into the construction of new orthodoxies, fully as constraining as any picture that held us captive.[26]

What we wish to suggest is that while it is true that every procedure, and not just critical reflection, is susceptible to misuse and abuse, just as Moran argues, reflection of the sort the proponent of the pruning view advocates is likely to be misused *systematically*. The reason, we would like to argue, is that reflection of the sort envisioned by proponents of the pruning view is not supplemented by a non-moralizing attempt to understand. It therefore delivers an evaluative judgment too quickly and an evaluation reached too early in the process of self-examination is likely to distort perception. Thus, Jacob the Bully, judging a desire to bully others to be unacceptable, is likely either to try to justify the desire so he doesn't disapprove of it anymore, or to convince himself he no longer has the desire.

Evaluative judgment which comes too early tends to block self-understanding by putting pressure on the agent to portray himself to himself as meeting his own evaluative standards. This appears to have been the case with the criminals studied by Fiske and Rai.

There is another problem here. Even if Jacob does recognize the desire to bully others by means of critical reflection of the sort championed by pruners, he never gets to find out just *why* he has that desire; moreover why he has agentially embraced it. From his point of view, the desire, if acknowledged at all, remains unexplained and perhaps, inexplicable.

Finally, critical reflection all by itself, even when it leads us to the right conclusion may utterly fail to produce a behavioral change. Why expect that it would produce changes? Perhaps, we think that it would, because we believe we have free will and can simply decide whether or not to act on a given motive. But there are psychological facts about humans that militate against this: willpower is a resource that can be depleted. Even with the best of intentions, Jacob is likely to sometimes fail to control his impulses and slip into his old ways, much the way a recent non-smoker may slip into her old addictive habit when too tired. No less importantly, even if Jacob does succeed in subduing the desire, he will have to expend emotional and cognitive resources in order to do so, thus leaving less ability for self-control in other domains or at any rate, fewer resources.[27]

Aristotelian variants of the pruning view may be thought immune to this problem, because on such variants, the desire can be expected to (hopefully) weaken after being repeatedly put under control. However, the Aristotelian strategy may or may not work. Jacob's desire to bully others may well *not* subside as a result of being controlled, at least not until his later years, when he, like many aggressive youths, can be expected to "mellow."

So what other alternative to "pruning" is there? This is the question we turn to now. We aver that there is an alternative way to proceed – examine the root causes of the need for chronic bullying, as one can do in psychoanalysis. The examination is initially non-moralizing: the goal is simply to

understand. This is what we would like to call the "understanding first" principle. Understanding must precede judgment, if we are to understand at all, particularly, when it comes to emotionally charged issues. Psychoanalysis has tools that facilitate the adoption of what we call the "understanding first" principle.[28] Patients in analysis gradually start to describe and/or enact the details with their analysts in the phenomenon known as transference. Usually, after a while, patients (and analysts) recognize a pattern both in their current bullying episodes (with the analyst and with others) and those from the past. Frequently, these patterns suggest that some early pathological identification has significantly contributed to the development of the negative trait. For instance, Jacob from our example may have had a harsh and violent father. He might have identified with his hostile aggressive parent, becoming like him in order not to identify *instead* with his father's downtrodden victim(s). This is a pathological identification. Just like more normal childhood identifications, pathological identifications happen smoothly, without fanfare, occurring and developing over time, and they do so mostly unconsciously. Perhaps most significantly, pathological identifications are set in motion early in life when the world seems to hold very few options, in this case – two: be a bully or be bullied.

In a successful analysis, these early pathological identifications become alive again in the relationship with the analyst, in the transference, a state in which often the patient (and sometimes the analyst too) had not initially been aware of re-playing the childhood situation. Interpretative interventions begin with demonstrating the repetition of these unconscious identifications. Once this has become manifest to both patient and analyst, the *agential* role of the patient in the formation of this identification can be observed and explored. Most patients can readily appreciate that they are the ones setting up the old patterns in the replays with the analyst through their own agential acts. The analyst has not initiated "bully or be bullied" – that is the patient's doing. In understanding their own agential role in the present transference situation with the analyst, patients can make changes. Minimally, when patients find themselves in a situation parallel to the original one, they become much

freer and more capable of breaking with old patterns. Maximally, patients can consider the idea that even as children, they did have choices although at the time, it may have felt as though they did not. The world was never actually binary – be a bully or be a victim – even though that is how it felt. Indeed, the child's *psychological reality* was such that he or she experienced only two alternatives, both not-good; but the analyst can and should point out that there were actually other potential possibilities – e.g., other adults with whom to identify, other models available.

We note however, that the analyst should also help the patient understand that his or her now troublesome identifications "made sense" at the time. An adult patient with the pathological trait of chronic bullying may have, for instance, identified with his abusive, aggressive father (a) in order to forge an alliance with him, (b) in the hopes of not identifying with the victims (as discussed above), and (c) to gain at least some gratification (albeit pathological) by bullying his younger siblings. Further, the analyst can explain that once the pathological identification was in place, the patient probably did experience the environment as more settled, more predictable, and more under the patient's own control. To the extent then that this self-feature – being a bully – did have a meliorating effect on some early psychological and socio-environmental problems, this negative trait would become stable, almost fixed. And this is the case particularly with traits developed by young children. After all, children are not yet the sort of scientists trained to test alternatives. Once a child has found something that appears to work well enough, given the nature of psychic economy, he or she tends to stick with it.

In the end phase of psychoanalytic treatment – after many cycles of rhythmic psychoanalytic investigations, examining the possible origins and explanations for "bad" traits without ever denying the patient's past and current agency in their development and maintenance, these un-preferred, regrettable, and unwanted negative traits, identifications, desires, etc. can finally be fully accepted as part of the – now evaluated and expanded – real self. This means such negative self-features are no

longer excluded, rationalized-away, externalized, or felt as alien. Patients admit and take agential responsibility for their bullying behaviors. Is the process of self-constitution then complete? While, in the final phases of an analysis much has been done, there is still work to do, post analysis, and thereafter. We will take this up specifically in the concluding Section 5.

But first one may wish to ask here how far we want to go with the suggestion that undesirable parts of the self are often parts of the proper self, that they may be agentially driven, and that understanding them is an important part of the task of self-constitution. In the beginning of this article, we expressed sympathy with conativists who argue that criteria of externality traditionally favored by strong agentialists – e.g., passivity and lack of reflective endorsement – are not necessarily marks of externality. But are there any good criteria? Aren't there aspects of the self that are truly "external" and that must be negated, as per the pruning view?

There are really two thoughts here: the first is that there may be truly external aspects of the self. The second is that the proper way to deal with those aspects *may* be to extrude them. We argued against the pruning view by casting doubt on the ways in which candidates for pruning are to be identified. However, the suggestion that correctly identified "external" parts must be negated and extruded sounds plausible: one may think that, much the way expulsion of foreign objects from the body is a necessary first step of the healing of a body, so likewise, taking aspects of the self that are external to the self – if there be any such – is a necessary part of the process of self-constitution. Perhaps, if there actually *are* truly external aspects of the self, pruned they must be. Are there? To this question we now turn.

4. How inclusive is the self?

Another way to pose the question is to ask whether there is a criterion we can use for distinguishing between self and non-self. The brief answer is that no one simple criterion can be given, since different

parts of the self can belong for different reasons. In addition, various elements of the self that *appear* to be "external" may, upon analysis, turn out not to be so. No less importantly, to the extent to which there are criteria, they cannot be directly action-guiding, because we cannot know whether a given criterion applies unless we engage in the process of self-constitution. The conclusion we'll draw from here will be that the theory of the self and the practice of self-constitution are inevitably intertwined, and self-constitution cannot be seen as a kind of application of the theory of the self.

4.1. Identification versus alienation[29]

For Frankfurt, the boundary between what properly belongs to the self and what doesn't is the boundary between what we identify with and what we feel alienated from. The mark of externality is the feeling of alienation. We previously rejected the pruning view, including Frankfurt's version of it, but perhaps, that was too quick. Frankfurt's recommendation that discordant elements of the self must be negated and extruded may be less vulnerable to our criticisms if the discordant elements are identified in the way Frankfurt himself suggests. Maybe, pruning sounds like a bad idea only if we focus on the wrong criteria of selfhood. What we wish to argue now is that Frankfurt's own criterion of externality will not make the pruning idea any more plausible. We'll meet him on his own turf and discuss a case that he himself considers to be a paradigm instance of externality.

Inexplicable Rage

> In the course of an animated but amiable enough conversation, a man's temper suddenly rushes up in him out of control. Although nothing has happened that makes his behavior readily intelligible, he begins to fling dishes, books, and crudely abusive language at his companion. Then his tantrum subsides, and he says, "I have no idea what triggered that bizarre spasm of emotion. The feelings just came over me from out of nowhere, and I couldn't help it. I wasn't myself. Please don't hold it against me.[30]

Frankfurt goes on to tell us that while these expressions may be "shabbily insincere," they may also be "genuinely descriptive." And also, "What the man says may appropriately convey his sense that the rise of passion represented in some sense an intrusion upon him."[31] But really, inexplicable rage may be inexplicable only on the surface. The man from Frankfurt's example may be caught in a trauma cycle. Triggers in his current environment may, due to their resemblance with triggers from his past, lead to fits of rage. Compare Frankfurt's inexplicable rage case with an example of unprovoked rage described by psychiatrist Janine Stevenson:

> There were a couple coming along for couple's therapy, because they were having problems in their marriage. They were in their forties, the children were fairly old and mostly had left home. They had a routine each evening – one night, one would cook while the other one set the table, and the following night the opposite would occur. So the first night, the wife was cooking, and the husband was setting the table. The wife looked across as the husband was setting the table and said, "Oh, don't use that cutlery. We'll use a different type of cutlery." He said, "OK," put the cutlery back and changed the cutlery. The following night the husband was cooking, and the wife was setting the table. The husband looked across and said, "Oh, don't use the old table cloth, use the new one." And his wife exploded. She got very angry and very frustrated and screamed at him and told him that he didn't care about her, didn't value her.

So far the case resembles Frankfurt's very closely. But here is how Stevenson interprets her case (and we are sympathetic):

> What was happening to this woman? She was obviously caught in a trauma system. A traumatic memory where the emotion is felt very strongly. She reacted to an environmental stimulus [resembling in some unconscious fashion, a past one] which triggered something like a flash bulb. The emotion was felt without the associated narrative.[32]

The idea here is that a traumatic memory is stored in the unconscious memory system (organized in a different fashion from the conscious memory system), and that external stimuli present in a person's current environment can revive the emotional reaction associated with the memory.

Let's now return to Frankfurt's case. It could be that the interlocutor of the angry man said something as innocent as, "Please don't forget to return this book," and the soon-to-be-enraged man, who, as it turns out, was once accused of stealing an item he'd simply forgotten to return, experiences a fit of rage. Why? Because the current stimulus triggers this emotion in so far as it resembles *psychologically* the old stimulus which had originally provoked it. The angry man from Frankfurt's example too may be caught in a trauma cycle and may remain trapped in it for as long as he does not engage in the type of deeper self-examination we advocate. He may, as in Frankfurt's vignette, recognize the complete lack of appropriateness of his response and declare it inexplicable. Or he may seek to justify it by finding reasons to be angry with his interlocutor. The former response is preferable for a number of reasons, but neither is optimal. The rage is only *seemingly* inexplicable; it is part of this person's self, and agentially driven, both in the original traumatic situation and at least partly in the current one.

Recall, however, that, as we acknowledged in the beginning, there are cases in which it seems intuitive to say that the agent is battling against something inside of him or her, just as Frankfurt would argue. In those cases, it is also plausible to say that whatever aspect of herself an agent is battling against is something external to her (she is the one waging the battle, and she is not the thing she is battling against). What are we to say of this?

While identification is certainly an important mechanism whose results can be suggestive, it is a fallible mechanism for distinguishing between what belongs to the self and what doesn't. We just argued that actions and attitudes from which we feel alienated may nonetheless be agentially driven. On the flip side, recall that OCPD patients identify with and endorse their irrational motives and the

behaviors which follow, yet arguably, identification does not make those attitudes "their own" in the relevant sense. In more ordinary cases, a person may mis-describe to herself her own motives and fail to identify with them for that reason. For instance, the person with a racist bias may believe that her sole reason for not inviting a candidate of color for an interview is the putative weakness of the candidate's file. Most importantly, perhaps, we must ask: identify with an action or an attitude *when*? The range of attitudes with which a person identifies expands over the course of psychoanalysis and can be expected to expand in any self-constitution practices driven by the principles we argue for. Thus, if an attitude, trait, or action are ruled out as "external" without deep self-examination from the get-go, the result would be a rushed self-pruning in which the baby will be thrown out with the bathwater.

4.2. Reasons responsiveness versus lack thereof

Another possible criterion for self inclusion vs. exclusion is reasons-responsiveness: it could be argued that regardless of whether or not we identify with an attitude or an action, all attitudes that are responsive to reasons belong to the self. If this criterion is correct, good candidates for pruning will be all aspects of the self not responsive to reasons.

This suggestion too has something going for it: intuitively conditions such as phobias may seem to be "external" to a person precisely because they are not reasons-responsive.[33] But reasons-responsiveness is considerably more plausible as a criterion of responsibility for action than as a criterion of identity. This is because there is a certain obduracy, recalcitrance, and lack of reasons-responsiveness to certain aspects of the self that are, for all that, decidedly an integral part of the self. Recall Margaret Gray whose love for Chris Bauldry seems unresponsive to reasons. And yet, the love is surely hers in a very deep way. This leads us to the next suggestion: cares. Perhaps, love is not "external" to a person because it expresses what a person deeply cares about.

4.3. Cares

Chandra Sripada, in the course of developing his version of what's been called a "Deep Self" view (which belongs, roughly, to the conativist family)[34] suggests that what defines the boundaries of a person's self are her *cares*: actions expressive of a person's cares are attributable to her. While Sripada is primarily concerned with the questions of attributability and responsibility, his account contains a theory of the self (hence, "Deep Self" view). If Sripada's theory is right, it will follow that the mark of externality is lack of caring.

Once again, the main suggestion sounds plausible. It may be thought, for instance, that being a philosopher is part of your identity while being a cartographer is not because, while you once made a map for a school assignment, the former but not the latter is expressive of what you care about. Plausible though as this suggestion may be, it must be resisted. Caring cannot serve as the sole criterion of selfhood. While there may be parts of the self that belong to the self because they express what we care about, other parts may belong to the self for other reasons. Thus, racial bias may be part of who a person is, even when it does not, in any plausible sense, express her *cares*: the bias may be a remnant of a person's upbringing, renounced by her, and inconsistent with everything she truly cares about.

Equally significantly, even if the criterion *were* correct, we would not be able to model self-constitution as an application of this criterion. After all, there could be things we don't care about but should – for instance, the achievement of long-term goals, not just the pursuit of temporary pleasures. There may also be things we care about but should not – for instance, various forms of systematic bias and/or personal envy frequently go unacknowledged for just this reason. So unless we think: a) that people are immune from self-serving false beliefs and b) that normative standards have nothing to do with self-constitution, we must reject the idea that self-constitution could be simply a matter of extruding what we don't care about.

We conclude from here that lack of caring does not stand up as an adequate mark of externality. What we argue now is that even to the extent to which a person's self is constituted by her cares, deep and initially non-moralizing self-examination of the sort we advocate would be necessary in order to identify what we *really* care about. Failing that, we are likely to misapply the criterion. Indeed, Sripada, in an earlier version of the published article we cited just above, misapplies it, or so we will maintain. We do not, by any means, wish to hold this against Sripada since, as we say, the example does not appear in the published version of the piece. Nonetheless, we wish to comment on it, because uncovering the source of the error will bring us to one of *our* central ideas. Consider the example in question:

Freezing on stage

A performing-arts aspirant freezes during her New York audition. Describing said performer as someone "…who desperately wants to be famous, so much so that when she is under pressure she is sometimes overwhelmed and freezes up" Sripada holds that: "…going to New York to audition expresses her self."[35] But he also claims: "Freezing on stage, however, does not express her self because caring for fame is not content congruent, in the goal-based sense of congruence, with her freezing."[36] Is the freezing "external" to the aspiring artist's self, as Sripada suggests here?

It is not clear. It may well be that while the performer has a very conscious (even desperate) desire for success, she is unconsciously ambivalent. She has an *unconscious* desire to hamper her own success and *unconscious* agential power to effect the freezing. While there could be many reasons for her ambivalence and self-defeating desires, these are not our concern here. What matters is that the artist's agential involvement in her freezing may not be obvious but may nonetheless be central to her cares[37], and thus, may be key to making sense of the freezing. This leads us to one of our most basic proposals: a viable account of self-constitution must recognize the role of *unconscious agency* in human actions. Without allowing unconscious agency as constitutive of this performer's otherwise puzzling

actions, they become almost inexplicable. This results in a very superficial understanding not only of her freezing, but far more fundamentally, of her deep self *itself*. Herein lies our answer to the question of how a seemingly inexplicable action issuing from an unintelligible (at least initially) "external" motive may nonetheless be central, internal, and agentially driven.

We suggested above that in the passage quoted, Sripada misapplies the criterion of selfhood he himself proposes. It is possible to dispute our point: perhaps the freezing – while it issues from unconscious desires and thus is in some sense agentially driven – is not expressive of what the aspirant cares about. In fact, one can go further in this direction and discredit the very idea of unconscious agency. In fairness, the idea is contested and certainly not universally endorsed. What evidence do we have for the existence of unconscious agency? We discuss this idea and its implications in a different project, but for present purposes, we wish to give two reasons in its support. First, unconscious desires may sometimes be gauged from our own reactions consequent to our actions. Thus, suppose Anette secretly looks at the documents on Joe's computer with the conscious belief that Joe appears to be hiding something along with the conscious hope (desire) that he is not. When she finds nothing that would reflect badly on Joe, she realizes that she feels disappointed. How is the disappointment to be explained? The plausible explanation is that Anette wanted to see Joe publicly disgraced – if all she had was the desire she consciously ascribed to herself, she should be glad to find out that her colleague is an honorable man. But she is not glad – she is disappointed.[38] Clearly, Anette's behavior is agentially driven, but it is largely driven by unconscious motives and so is largely a result of unconscious agency.

Second, while the label "unconscious agency" may be widely resisted, other labels in the vicinity are widely accepted, for instance, implicit bias. Thus, barring neurological problems or other unusual extenuating circumstances, the person with an implicit racial bias *is* generally regarded as racist to some extent. Further, the discriminatory actions of a person with an unacknowledged racial bias are taken to be agentially driven. But if implicit racism is an expression of a person's agency, so are a

host of other unacknowledged propensities, including those not morally charged, such as the self-sabotaging desires of the aspiring artist.

Our "barring neurological problems" qualification may point to a third possible criterion: sanity. Perhaps, everything that issues from a sane self belongs to the self. We turn to this possibility presently.

4.4. Sanity

Taking our cue from Sarah Buss, who argues for sanity as the mark of autonomous agency, we consider the possibility that whatever issues from a sane self belongs to the self.[39] On this view, the boundary between self and non-self is that between sanity and insanity.

Once again, the suggestion sounds plausible. We do sometimes take lack of sanity to be a mark of externality, as when we say to someone, "It's not because he doesn't care about you that he cannot recall your name, it's because he has Alzheimer's," alleging that the name-forgetting is to be attributed not to the person who forgets but to his disease.

But this suggestion too is not without serious difficulties. Arguably, people with narcissistic and antisocial personality disorder are not, strictly speaking, fully sane, but the narcissistic and antisocial traits they possess are very much a part of their selves. Moreover, people may have had an agential role to play in their own psychiatric problems, for instance, a person can develop a Dissociative Identity Disorder (DID) in an attempt to protect herself from painful experiences. This is not to suggest that the DID patient actively chooses to be pathologically dissociated, but it does mean that neither we nor she can truly make sense of her condition without appealing to her agency and saying that she was "trying to protect herself." Similar considerations apply to the cases of people with angry outbursts we discussed earlier: seemingly inexplicable angry outbursts are, arguably, not compatible with complete sanity either, but they may be agentially driven.

Even if sanity *were* the correct criterion, however, here too construing self-constitution as a mere application of the criterion, then extruding whatever doesn't match, will not work. Why not? Because we may not be able to tell whether particular actions or attitudes are compatible with sanity without deeper self-examination. Is a person's anxiety pathological or a normal response to her circumstances? And what about her angry outburst? In order to be able to answer such questions, we may have to engage in precisely the type of deep and initially nonmoralizing self-examination we have been advocating. We will need to discover whether, for example, we are truly responding to the stimuli in our current environment, or merely seeming to address them consciously, while unconsciously reacting to past stimuli that resemble (in some fashion) the current ones.

We have shown all the criteria discussed so far to be insufficiently inclusive for proper self constitution. Perhaps, we can try to be as inclusive as possible, evaluating the view that the skin is the boundary between self and non-self.[40]

4.5. The skin as a boundary

As with the other candidates, this idea also has some appeal. If you touch my skin, I will say you have touched *me*, and if you burn by skin, I will say you have burned *me*. If you touch or burn my coat, by contrast, I will say you have touched or burned my property, not me.

The obvious problem with this way of drawing the boundary is that much of what happens on the inside of the skin, while it is attributable to a person's organism, does not, strictly speaking, appear to be a part of the person. For instance, a person's digestive processes are not attributable to her as an agent. Actions and attitudes that result directly from a neurological problem do not seem to be attributable to her in the relevant sense either. It is likely that the recognition of this point, in particular, has been instrumental in proposing criteria we've discussed such as caring or identification for self vs. non-self ascription. Though we argued that those criteria fail, we agree that a criterion more restrictive than skin boundary is needed. What more restrictive criterion could we adopt?

One possibility is to exclude those events or features of personality that have a physiological rather than a psychological basis. The first problem with this suggestion is that it is incomplete: it does not tell us how to distinguish psychological from physiological causes. But suppose we choose to fall back on intuitions here and said that there is simply an intuitive difference between physiological and psychological causes. Still, there is a second problem: a person's *physiological* processes may have a *psychological* cause. For instance, stress – which can have a purely psychological cause – can increase the acidity in the stomach and cause indigestion or other somatic symptoms.[41] A person in this case has an agential role to play in her own indigestion, in a way in which a person whose indigestion is caused by food she cannot process does not have an agential role to play. We must therefore conclude, once again, that even if this criterion were correct, we would not know how to apply it without engaging in something like the process of deeper self-examination we champion.

Note also that the skin criterion, while in one sense too promiscuous and hence in need of additional restrictions may, in another sense, be too restrictive. Consider, for instance, the fact that many events that occur on the outside of a person's skin become a part of a person's life history and to that extent, a part of her self. A person becomes a grandmother on the day her first grandchild is born and a Nobel Prize winner on the day the Swedish Academy bestows on her the prestigious award. Being a grandmother and a Nobel Prize winner now becomes, in both her eyes and in those of others, part of her identity. Perhaps, one can say that it is only because a person cares about these things or identifies with them that we associate her with them. But this is not right. A grandmother who doesn't care is an uncaring grandmother, and that's part of who she is as well (though she may not want to think of herself in that way), and a Nobel Prize winner who doesn't care about the Nobel Prize is just that – a Nobel Prize winner who, unlike others, is not taken by her own success. Similarly, a grandmother who does not identify as a grandmother, say because she believes she is way too young to be a grandmother is nonetheless a grandmother too. Again, a Nobel Prize winner with an impostor

syndrome, who does not identify as a Nobel Prize winner and has the irrational belief others will discover that she does not deserve the Nobel award and take it from her is nonetheless a Nobel Prize winner, just one with an impostor syndrome. More generally, societal standards of identity appear relevant to what we consider a part of a person's self. This is true even when a person rejects those standards or when they clash with what she cares about or identifies with.

We began this section by asking whether our criticisms of the pruning view do not derive their force from an untenable criterion of externality. The thought was that perhaps, if we adopted the correct criterion of selfhood, elements alien to that self would emerge. We then examined five different criteria for distinguishing what belongs to the self from what doesn't, and we found problems with each. We also argued, however, that there is something intuitively right about each criterion. Each captures some part of the truth, just not the whole truth. The first conclusion we wish to draw from here is that the self contains disparate elements that may belong to it for different reasons: some parts of the self may belong because we identify with them consciously; others – because they issue from our unconscious agency, although we do not consciously identify with them; still others may belong because they express what we care about, whether passionately, passively, consciously or even unconsciously.

Perhaps more importantly, we argued that even to the extent to which each of these five criteria captures something intuitive, it is unclear how each is to be applied, i.e., what falls within the criterion's boundaries and what doesn't. The second conclusion we wish to draw is that our criticisms of the pruning view derive their force from a weakness in the view. On the pruning view, one must begin by identifying the boundaries of the proper self and then seek to disarm and cut out the elements on the outer side of the boundary. This strategy, however, is unlikely to work. On the alternative we advance here, discovering via self-examination, elements of the self that are agentially driven – even negative elements – contributes much to determining and constituting the self. Thus, this self-

evaluative process itself forms the backbone of self-constitution. The purpose of self-constitution, we contend, is not simply to place unwanted desires and motives outside the boundary of the self, as is sometimes supposed; it is as much allowing an enlarged notion of self, first discovering, only thereafter determining where the boundary shall lie.[42] Indeed, one may discover that negative aspects of the self, owned as agentially driven, but arising from an earlier time, are no longer motivated (as was the case for Jacob, now an ex-bully), and thereby no longer part of the self proper as constituted.

5. Conclusions

Let us finally take stock. In this article, we discussed the self and self-constitution. We began by criticizing what we termed the "pruning view" of self-constitution and its central account of the self. We next suggested that "conativists" who propose accounts of the self more inclusive than those underlying the different versions of the pruning view are on the right track. Although we later expressed doubts regarding conativists' attempts to give criteria for distinguishing what belongs to the self from what doesn't – for instance, we showed that the use of either cares or sanity was problematic – we claimed that in fact, the *main* problem with conativist theories was that they do not offer any viable accounts of self-constitution.

On the constructive side, we proposed an account of self-constitution that incorporates key insights from psychoanalysis. We find our views largely in sync with those of Jonathan Lear. Lear endorses, as we do, transference as a particularly potent process in promoting "understanding first." Further for Lear, no less than for us, the role of unconscious agency and motivations are central to one's self and therefore must be taken into any account of self-constitution. As Lear notes, while we typically think that conscious organizing principles impose their form on the unorganized matter of unconsciousness, the reverse is true as well: unconscious principles can use conscious processes as *their matter*.[43] They can shape reflection, motivate belief-formation, distort self-perception and do a

host of other things that militate against self-constitution. We must, as it were, reach a truce with those unconscious processes and contents, first recognize and acknowledge them as our own, then reconcile with them, and enlist them in the enterprise of constituting ourselves. The view we offer in this article also has resonances with Charles Taylor's view of radical evaluation, and accommodates the inclusivity of (the most inclusive among) conativist approaches while also offering a viable path to the goal of inner integration toward relative harmony.[44] Whereas on the pruning view, agential self-constitution requires actively extruding one's negative self elements and making them alien, we hold just the opposite, namely, that to understand and fully *be* oneself, one must embrace and take responsibility for the negative aspects of the self, no less than the positive.

Psychoanalysis, the process from which we developed our view of self-constitution, often begins with patients reluctantly acknowledging that undesirable features or parts of themselves are in fact parts of themselves. These can include (a) undesirable desires, for instance wanting illicit substances or forbidden sexual partners; (b) regrettable character traits such as bullying or submissiveness; and (c) unacceptable irrational fears. No work in psychoanalysis can take place until patients *stop* externalizing and *stop* rejecting these undesirable self-parts. Thus patients who externalize and claim that such behaviors, desires, and emotions, while regrettably subjectively experienced, are nonetheless alien and in no way agentially driven, must first come to terms with these unsavory negative portions of themselves. Once expansion of the self contra the pruning view has been accomplished, the proceeding stages of psychoanalysis can begin.

Next, and still very early in the analysis, patients can choose not to act on the desires and motivations that follow from negative aspects of the self. This sort of second-level agential choice can stop or at least minimize actual damaging life events, even as these aspects are not yet well understood.

But analysis does not end here. Once the patient can fully *embrace* the fact of his or her *agential* role in constituting and maintaining the seemingly peripheral and even external parts-of-self – the

negative pathological elements as well as the preferred aspects – the patient can go forward to make different (better) future agential choices. This can mitigate the effect of the undesirable elements, if not entirely omitting them. In fact, one of us (initials withheld for blind review) would endorse the following comment from a person whose name we delete for purpose of blind review, who states, "…in order to get rid of any part of yourself, you have to have been involved in the creation of that part. If you have not been involved in the creation of a given attitude, then you can't undo it."[45] In any case, these future choices can look much more like the type of second-order volitions that are central to the pruning view, especially that of the later Frankfurt, with endorsement and identification constituting one's will and one's self, toward exercising autonomy and choice, as a whole person. Psychoanalysis, then, does not, and in fact cannot, in itself achieve personhood constitution – it functions only to remove pre-existing formidable neurotic barriers. Yet in this arduous if modest way, psychoanalysis frees each patient to truly choose what sort of person to become.

---

[1] The idea of perfect integration in the realm of practical reason may seem as untenable as the idea of perfect coherence in the domain of belief, or theoretical reason.

[2] See Plato, *Republic*, translated by G. M. A. Grube in *Plato: Complete Works*, edited by John Cooper (Indianapolis, Hackett Publishing Company, 1997); Immanuel Kant, *Critique of Practical Reason*, translated and edited by Mary Gregor (Cambridge, UK: Cambridge University Press, 1997); his *Groundwork of the Metaphysics of Morals*, translated and edited by Mary Gregor (Cambridge, UK: Cambridge University Press, 1998); and his *Religion within the Limits of Reason Alone*, translated and edited by Theodore Greene and Hoyt Hudson (New York, NY: Harpertorch Books, 1960); Joseph Butler, *Fifteen Sermons* (Cambridge: Hilliard and Brown; Boston: Hilliard, Gray, Little, and Wilkins, 1827), esp. Sermon II, "Upon the Natural Supremacy of Conscience"; Christine Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press 1996) and her "Self-Constitution in the Ethics of Plato and Kant," *Journal of Ethics* 3 (1999): 1-29; Tamar Shapiro, "What is a Child," *Ethics* 109 (1999): 715-38; Harry Frankfurt, "Identification and Externality," in *The Importance of What We Care About* (New York, NY: Cambridge University Press, 1998), 58-68.

[3] David Velleman, "The Self as Narrator" in *Self to Self: Selected Essays* (Cambridge, MA: Harvard University Press, 2006), 203-223, 203.

[4] John Doris, in his recent book *Talking to Our Selves* (Oxford, UK: Oxford University Press, 2015), 17 adopts a similar strategy. His goal is to attacks a view he calls "reflectivism." In addressing his task, he opts not to discuss particular versions of the view held by different philosophers. Instead, Doris constructs a composite image out of a number of reflectivist type accounts, and his criticisms are directed against *that picture*.

We here, much like Doris, intend to highlight the predominant core ideas of all of those we've termed "pruners;" views that have been variously described as "the rational self," "the endorsed self," and "the true self" rather than focusing on the details of any given version of the view. As Doris notes in a parallel move, this strategy is note without risks for it is: "…rather like drawing a composite face that looks a little like many faces, but not a lot like any particular face…" Understandably, some of the philosophers whose accounts fall under the general umbrella are likely to object by saying that their views have been partly misrepresented. But Doris goes on, "Nevertheless, I'm betting that the face I depict is easily recognizable: the commitments here depicted in composite strongly resemble commitments actually held," 17. We

believe, similarly, that the pruning view we outline is recognizable and sufficiently resembles "views actually held." And as we prefer to avoid spending too much time and detail on the details of its various versions, we will work mainly with the general picture as we state it although in Section 4, we will say more about Frankfurt's version specifically, because Frankfurt makes a suggestion that ought to be examined more closely.

[5] We take it that rationality and reflection are not the same thing. Rationality has to do mainly with reasons-responsiveness. But a creature can be reasons-responsive – and to that extent rational – without being reflective–(arguably, some non-human animals possess rationality but not reflective capacities). On the flip side, reflection can co-exist with a good deal of irrationality – an irrational and even disordered mind can be reflective, for instance, an excessive worrier or a paranoid person can be reflective (even hyper-reflective).

[6] Frankfurt, "Identification and Externality," 67.

[7] Gary Watson in "Free Agency," *The Journal of Philosophy* 72 (1975): 205-220, 217-219 raises a similar objection, arguing that Frankfurtian second-order desires cannot do the work Frankfurt wants them to, because a second-order desire, being merely another desire, has no special status. In addition, as Frankfurt himself acknowledges, one could raise an infinite regress objection here and ask, "Why stop at the second-order level?" Why not form third-order desires with regard to our second-order desires? In response to such objections, Frankfurt has modified his view, arguing that the infinite regress can be stopped and the possible arbitrariness of one's second order desires countered by making a "decisive commitment" to a certain course of action. See his "The Faintest Passion," *Proceedings and Addresses of the American Philosophical Association* 66 (1992): 5-16.

[8] She writes, "in order for you to achieve this unity, your actions must be in accord with morality. Integrity in the moral sense and for persons integrity in the metaphysical sense are one and the same thing." See her "Self-Constitution and Irony," in Jonathan Lear, *A Case for Irony* (Cambridge, MA: Harvard University Press, 2011), 75-83, 76.

[9] Christine Korsgaard, "Self-Constitution in the Ethics of Plato and Kant," 12.

[10] We say "on the face of it," because it is certainly possible for a person to identify *more* with one of her first-order desires than she does with another, and others may identify her so as well, for instance, the agent from the example may see herself as a big sleeper but not as a big eater, and if that's the case, it may be plausible to say that to continue sleeping would be "her style" while to get up underslept in order to have breakfast would be "unlike her."

[11] Note that while Frankfurt puts forth a version of the pruning view, his position actually occupies a sort of middle-ground since his view combines broadly conativist sympathies with a hierarchical model of the psyche similar in structure to those of cognitivist proponents of the pruning view such as Korsgaard.

[12] See, for instance, Nomy Arpaly, *Unprincipled Virtue: An Inquiry into Moral Agency* (Oxford: Oxford University Press, 2003) and Nomy Arpaly & Timothy Schroeder, "Praise, Blame and the Whole Self," *Philosophical Studies*, 93(1999): 161-88; Sarah Buss, "Autonomous Action: Self-Determination in the Passive Mode," *Ethics*, 122 (2012): 647–691; Terence Penelhum, "The Importance of Self-Identity," *Journal of Philosophy* 68 (1971): 667-678; Thomas Scanlon, *What We Owe Each Other* (Cambridge, MA: Harvard University Press, 1999); David Shoemaker, "Caring, Identification, and Agency," *Ethics* 114 (2003): 88–118; Angela Smith, "Responsibility for Attitudes: Activity and Passivity in Mental Life," *Ethics* 115 (2005): 236–271; Chandra Sripada, "Self-Expression: A Deep Self Theory of Moral Responsibility," *Philosophical Studies* (in press). Available at: http://sites.lsa.umich.edu/sripada/by-theme/deep-self/. Last retrieved December 21, 2015. Manuel Vargas argues, in an ecumenical spirit, that there is really no fundamental distinction between "real selves" approaches, on the one hand, and "reasons" approaches, on the other. See his "Reasons and Real Selves," *Philosophy* (2009), paper 5, available at: http://repository.usfca.edu/cgi/viewcontent.cgi?article=1003&context=phil. Last retrieved December 21, 2015.

[13] Susan Wolf in *Freedom within Reason* (New York, NY: Oxford University Press, 1993), has endorsed a type of asymmetry: we are praiseworthy for unthinking heroism but not blameworthy for actions that issue from completely unchosen character traits.

[14] This is the finding of David Rand and Ziv Epstein, who studied a number of Carnegie Hero Medal Recipients. See their "Risking Your Life without a Second Thought: Intuitive Decision-Making and Extreme Altruism," *PLoS ONE* 9 (2014): e109687, doi:10.1371/journal.pone.0109687. Note also that Ted Huston et al., "Bystander Intervention into Crime: A Study of Naturally Occurring Episodes," *Social Psychology Quarterly* 44 (1981): 14-23, find that people who intervene to help in criminal episodes such as armed robberies or muggings tend to have more emergency training and to think of themselves as physically stronger compared to non-interveners. Huston et al. conclude that interveners tend to be motivated not by strong humanitarian motives but by a sense of capability. This suggests that the actions of bystanders may be only partly automatic, since bystanders tend to intervene when they believe they will succeed.

[15] John McDowell, for instance, suggests in this connection that virtue cannot be "the outcome of a blind, non-rational habit or instinct, like the courageous behavior – so called only by courtesy – of a lioness defending her cubs." See his "Virtue and Reason" in *Virtue Ethics*, edited by Roger Crisp and Michael Slote (New York, NY: Oxford University Press, 1997), 142. Tamar Shapiro in "What is a Child?" offers a similar view. By contrast, Neil Sinhababu has recently argued that the proponent of the desire-based account of virtue should accept the consequence that non-human animals can possess virtues of character. See his "Virtue, Desire, and Silencing Reasons," forthcoming.

[16] A person whose name we omit for purpose of blind review suggests that Korsgaard's response here cannot resist the conativist criticism: "…it remains unclear how you can be responsible for defective actions if these, by definition, arise out of aspects of yourself that are not part of your self proper."
[17] Penelhum, "The Importance of Self-Identity," 674.
[18] Thanks to a person whose name we remove for purpose of blind review, who points out: "The idea that the desires we might wish to renounce might well be part of the self, and that as such we might not be able to renounce them…is simply a cautionary note about the possibility that we might misidentify whether particular elements are part of the self or aren't."
[19] Rebecca West, *The Return of the Soldier* (Rockville, MD: Wildside Press, 2009).
[20] This answer does not apply to Frankfurt, whose version of the pruning view is one of those we draw on here: as pointed out by Frankfurt's critics, on (early) Frankfurt's view, the second-order volitions relevant to self-constitution are simply desires. Thus, it is not critical reflection which picks out candidates for pruning but meta-desires. This may lead one to think that the objection does not apply to Frankfurt. In some sense that's right, but Frankfurt's account faces even more serious problems. While the standard process of critical reflection is, as we shortly argue, systematically flawed and likely to frequently get things wrong, Frankfurt's conception does not even contain a suggestion about what it would mean to get things right. See also footnote 7 above.
[21] Rebecca West, *The Return of the Soldier*, 68.
[22] Alan Fiske and Tage Rai, *Virtuous Violence* (Cambridge, UK: Cambridge University Press, 2014). Note that on Fiske and Rai's view, violent criminals really do have moral motives and are not just giving post hoc rationalizations of their behavior. We agree that the criminals studied may, indeed, act from such motives as "perceived breach in hierarchy," and in this sense, the justifications they give of their actions are not merely *post hoc* (since they could have motivated the actions under discussion). But we disagree that they act from moral motives: on our view, distorted reflective processes lead them to (falsely) *believe* they do.
[23] "The 'Breaking Bad' Syndrome? UCLA Anthropologist Exposes the Moral Side of Violence," *UCLA Newsroom*, December 22, 2014. Available at: http://newsroom.ucla.edu/releases/breaking-bad-syndrome-UCLA-anthropologist-exposes-moral-side-violence. Last retrieved December 22, 2015.
[24] This is the difference between OCD and OCPD: while OCD patients typically perceive their own rituals and idiosyncratic behaviors as irrational and unjustified, for OCPD patients, the behaviors are more fully integrated with the self. See *DSM* 5th Edition (Washington, DC: American Psychiatric Publishing, 2013), 235-242 & 678-682.
[25] See his *A Case for Irony*. Doris in *Talking to Ourselves* makes this point as well.
[26] Richard Moran, "Psychoanalysis and the Limits of Reflection" in Jonathan Lear, *A Case for Irony*, 103-114, 108.
[27] A phenomenon known as "ego-depletion." See Roy Baumeister et al., "Ego-Depletion: Is the Active Self a Limited Resource," *Journal of Personality and Social Psychology* 74 (1998): 1252-65; Matthew Gaillott et al., "Self-Control Relies on Glucose as a Limited Energy Source: Willpower is More than a Metaphor," *Journal of Personality and Social Psychology* 92 (2007): 325–336; and Haggar et al., "Ego-Depletion and the Strength Model of Self-Control: A Meta-Analysis," *Psychological Bulletin* 136 (2010): 495-525. We note that the ego-depletion idea has been challenged. See, for instance, John Lurquin et al., "No Evidence of the Ego-Depletion Effect Across Task Characteristics and Individual Differences: a Pre-Registered Study," *PLoS ONE* 11 (2016) and John Lurquin and Akira Miyake, "Challanges to the Ego-Depletion Research Go Beyond the Replication Crisis: A Need for Tackling the Conceptual Crisis," *Frontiers in Psychology* 8 (2017): 568. Whether or not ego-depletion is an empirically robust phenomenon, however, the general point remains that spending resources on counteracting unwanted desires is non-ideal.
[28] One of us (initials deleted for the sake of blind review) believes that while psychoanalysis may be helpful, it isn't strictly speaking necessary in order for a person to be able to put our model to practical use: there are various paths to self-knowledge, of which psychoanalysis is but one.
[29] Identification and alienation may converge in certain cases. In an interesting recent paper, Suzy Killmister argues that sometimes, we may identify with an action precisely because we feel alienated from it, a phenomenon she labels "authentic alienation." See Suzy Killmister, "The Woody Allen Puzzle: How 'Authentic Alienation' Complicates Autonomy," *Nous* 49 (2015): 729-747.
[30] Frankfurt, "Identification and Externality," 63.
[31] Ibid.
[32] Janine Stevenson, "Dissociation in Borderline Personality Disorder," available at: http://www.psychevisual.com/Video_by_Janine_Stevenson_on_Dissociation_in_borderline_personality_disorder_BPD.html.
[33] Although phobias are not reason-responsive—one cannot merely inform a spider phobic that a nearby Huntsman spider is harmless and expect fear to subside—phobias to spiders do develop for reasons (in a purely motivational – not a justificatory sense – of "reason").
[34] Sripada, "Self-Expression."

[35] Chandra Sripada, "Self-Expression: A Deep Self Theory of Moral Responsibility" (pre-publication version), 21.
[36] Ibid.
[37] It seems that Sripada himself regards the freezing as central. Note his comment quoted above. She "…desperately wants to be famous so much so that when she is under pressure she is sometimes overwhelmed and freezes up."
[38] One of us (initials omitted for blind review) has used examples of this sort in a previously published paper on what it means to act for reasons if we are wrong about our reasons. See (Author)
[39] We note that Buss in "Autonomous Action," seeks not a criterion of selfhood but a criterion of autonomous agency and responsibility. She wants a criterion which would allow us to make sense of agency exercised passively. She proposes sanity: sane people are responsible for what they do passively, through lapses of attention, for instance (a phenomenon Buss describes as "autonomous agency in the passive mode"). We consider the question whether her criterion of responsibility can also be seen as a criterion of what belongs to the self.
[40] There is a parallel question one might ask: which *actions* are externally as opposed to internally caused? For instance, if Dexter bribes Ally, are Ally's actions caused by the bribe or by the fact Ally is bribable? This is a question discussed by John Sabini, Michael Siepmann, and Julia Stein in "The Really Fundamental Attribution Error in Social Psychological Research," *Psychological Inquiry* 12 (2001): 1-15. These three authors reject the skin as a boundary between internal and external causes, but they claim that the criterion has been endorsed by others, hence that they are not creating a straw man by taking it up. In this connection, they quote from Gilbert and Malone's discussion of correspondence bias, "Although these theories differ in both focus and detail, each is grounded in a common metaphor that construes the human skin as a special boundary that separates one set of 'causal forces' from another. On the sunny side of the epidermis are the external or situational forces that press inward on the person, and on the meaty side are the internal or personal forces that exert pressure outward," Daniel Gilbert and Patrick Malone, "The Correspondence Bias," *Psychological Bulletin* 117 (1995): 21-38, 21 quoted by Sabini, Siepmann, and Stein, "The Really Fundamental Attribution Error," 8-9.
[41] In fact, because somatic symptoms are easier to detect and recognize as a problem, people often seek therapy because they have such symptoms.
[42] At times in all of us, there is the motivation to reject undesirable aspects of the self. In Section 5-Conclusions, we will discuss how on our view, perhaps uniquely, one can resist this tendency, and the advantages in so doing.
[43] Lear, *A Case for Irony*, 90.
[44] Charles Taylor, *Sources of the Self: The Making of Modern Identity* (Cambridge, MA: Harvard University Press, 1989).
[45] One of us (initials deleted for blind review) disagrees with this point on the following ground: people can generally change parts of themselves that are due to factors utterly beyond their control. Thus, if someone injected Betty with a chemical that makes Betty distrust race X, it certainly doesn't follow from here that there is nothing Betty can do to change her attitude. Similarly, a person with a genetic predisposition to get angry can do something about that disposition as well (not only about the behavior that results from it), and a person brain-washed to believe in the justice of violent means can come to reconsider this. The other author (initials withheld) thinks that external control cases, as well as cases involving erratic behavior owing to brain tumors, etc., are exceptions. We both believe that when it comes to genetic predispositions, the behavior issuing from the disposition must be licensed by the agent, at least tacitly. Both of us agree also that the disposition itself can come about as a result of factors beyond the agent's control, such as genes. But one author thinks that nonetheless, it may be possible to change the disposition itself – not just the behavior that issues from it – by using various techniques such as meditation or aversive conditioning. We both thank someone whose name we omit for purpose of blind review for the input.