Jada Twedt Strabbing
Fordham University
Penultimate draft. Paper in *Philosophy and Phenomenological Research*, 2016

Responsibility and Judgment

Abstract: I focus on the type of responsibility that an agent has for actions that express his practical identity, making it appropriate to evaluate him on the basis of those actions. This kind of responsibility is often called *attributability*. In this paper, I argue for a novel view of attributability – the Judgment Responsiveness View (JRV). According to the JRV, an agent is attributability responsible for an action A if and only if A results from either 1) his responding to his judgments about the (normative) reasons that he has in favor of doing A by doing A or 2) his failing to exercise his capacity to respond to his judgments about the (normative) reasons that he has against doing A by not doing A. The JRV diverges from other views of attributability for actions in two significant respects. First, it is not reasonably thought of as a "deep self view." According to deep self views, attributable actions are actions that express deep features of the agent, such as his fundamental values, cares, or commitments. As I show, thinking in terms of the deep self is too narrow for attributability. Second, unlike other views, the JRV claims – via condition 2) – that we can be attributionally responsible for actions that result from *failing* to exercise the attributability-relevant capacity to avoid them. My argument for the JRV thus shows that attributability is a broader and richer conception of responsibility than has been previously thought.

What makes an action attributable to an agent, such that he is attributionally responsible for it? Following Gary Watson, I take attributable actions to be those actions that express what an agent is like practically – i.e., that express his practical identity.[1] The idea is that when an action expresses an agent's practical identity, the action is attributable to him *as an agent*, making him responsible for it in a deeper sense than merely being its cause. Hence, in asking what makes an agent attributionally responsible for an action, we are asking: what does it take for an action to express an agent's practical

[1] Gary Watson. "Two Faces of Responsibility," in his *Agency and Answerability*, Oxford University Press, 2004. (Original work published 1996).

identity? To give an account of attributability, we should therefore determine a) what constitutes an agent's practical identity and b) what it means for an action to express that practical identity.

In this paper, I argue for a novel view of attributability – the Judgment Responsiveness View (JRV). According to the JRV, an agent's practical identity is constituted by his judgments about normative reasons, and his action A expresses those judgments – thus making him attributionally responsible for A – just in case A results from either his responding to his judgments about reasons in favor of doing A by doing A or his failing to exercise his capacity to respond to his judgments about reasons against doing A by not doing A. Although I think that the JRV is the right view of attributability generally, I here argue for it in the context of actions of moral significance. Thus I argue that the JRV is the right view of moral attributability, telling us which actions express an agent's moral identity. This is a reasonable restriction given that the literature on attributability primarily focuses on moral attributability as a distinct type of moral responsibility. Plus, the right view of moral attributability plausibly extends to attributability generally. (From this point, I will generally say "attributability" rather than "moral attributability.")

The JRV diverges from other views of attributability for actions in two significant respects. First, it is not reasonably considered a "deep self view." According to deep self views, attributable actions just are actions that express the "deep self"– i.e., actions that express deep features of the agent such as his fundamental values, cares, or commitments. As I show, thinking in terms of the deep self is too restrictive for an adequate account of attributability. Second, unlike other views, the JRV claims that we can be attributionally

responsible for actions that result from our *failure* to exercise the attributability-relevant capacity to avoid them. As a result, I demonstrate, only the JRV accounts for attributability for weakness of will.

My argument proceeds as follows. In Section 1, I illuminate the nature of attributability and bring out its connection to the appropriateness of appraisal in terms of virtues and vices, broadly construed to include any appraisal of the agent that goes beyond appraising him merely for the moral quality of his action (for actions of moral significance). In Section 2, I use this connection to argue that attributable actions are actions that are "up to the agent," in the sense that he has control over the fact that he performs them. In Section 3, I present and clarify the JRV, which cashes out the idea of agential control. In Section 4, I use the connection between attributability and the appropriateness of appraisal in terms of virtues and vices broadly construed to argue for the JRV. Specifically, I show that the JRV gets the right results for the right reasons in the hard cases of compulsion, weakness of will, and brainwashing. My argument for the JRV further reinforces the idea that attributable actions are actions that are up to the agent. Finally, in Section 5, I demonstrate that the JRV is a significant departure from the other views of attributability for actions in the above-mentioned ways.

1. Attributability and Appraisal in terms of Virtues and Vices

1.1. Understanding Attributability: An Example

Imagine two drug addicts. The first is a willing addict whose "willingness" to take the drug leads him to take it. By this, I mean that he is motivated to take the drug by

whatever represents the fact that he is willing – whether that is his values, cares, or something else. For example, he may be motivated to take the drug by the fact that he values the experience of being high. Of course, as an addict, he also has a compulsive desire for the drug. We can explain this by saying that his compulsive desire either a) also motivates him to take the drug, such that he is overdetermined to take it, or b) acts as a failsafe device, such that it does not motivate him to take the drug but would motivate him if he were not already motivated by whatever represents his willingness. In contrast to this willing addict, imagine the unwilling addict described by Harry Frankfurt, who "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."[2] Admittedly, this is an extreme unwilling addict, and few if any real-life addicts fit this description. But for my purposes, it is enough that we can imagine such a case.

Of these addicts, only the willing addict is attributionally responsible for taking the drug. His drug use expresses his practical identity, which is clear from the fact that his willingness to take the drug motivates his taking it.[3] The unwilling addict's drug use, however, does not express his practical identity because he is a victim of his compulsive desire; he tries to resist the desire but is unable to do so.[4]

These addicts also illustrate the kind of freedom required for attributability. As the willing addict shows, the freedom required for attributability is not the freedom to do

[2] Harry Frankfurt. "Freedom of the Will and the Concept of a Person," in Gary Watson (ed.) *Free Will*, Oxford University Press, 2003, p. 328. (Original work published 1971)
[3] Below I consider a willing addict whose drug use is motivated only by his compulsive desire.
[4] Here I ignore complications that arise from how these addicts became addicted. The unwilling addict may be attributionally responsible for that, and the willing addict may not be. My point is that, once they are addicted, only the willing addict is attributionally responsible for *continuing* to use the drug.

otherwise. He lacks this freedom but is still attributionally responsible for taking the drug. Instead, the freedom required for attributability is simply the freedom to express one's practical identity in action. The willing addict clearly has this freedom, since his addiction does not prevent his willingness to take the drug from motivating him to take it. However, the unwilling addict lacks this freedom because his addiction is a psychological obstacle to his expressing his practical identity in action.

Note that the freedom to express one's practical identity in action does not require normative competence, such as the capacity to recognize and respond to moral reasons. After all, due to his addiction, the willing addict lacks the capacity to respond to moral reasons not to take the drug, but his drug use still expresses his practical identity. Further, a psychopath is often thought to lack the capacity to recognize moral reasons altogether,[5] but he can still express that he values money over others' wellbeing when he empties his lover's bank account.

Because attributability requires neither the ability to do otherwise nor normative competence, I follow Watson in taking attributability to be a type of responsibility distinct from accountability, as accountability is often thought to require one of these. Yet you need not accept this distinction to accept my view of attributability. For example, if you take responsibility to be a unified notion and understand it as I understand attributability, you could think of my view of attributability as simply a view of responsibility.[6]

[5] For example, Blair et al. argue that psychopathy is marked by emotional impairments that impair moral reasoning. See James Blair, Derek Mitchell, and Karina Blair 2005. *The Psychopath: Emotion and the Brain*. Wiley-Blackwell, 2005.

[6] Angela Smith, for example, is in this category. See Angela Smith. "Attributability, Answerability, and Accountability: In Defense of a Unified Account." *Ethics*, Vol. 122, No. 3 (April 2012), pp. 575-589.

1.2. The Connection to Virtues and Vices

As Watson argues, it is appropriate to evaluate an agent *as an agent* on the basis of attributable actions precisely because these actions express her practical identity.[7] For example, a scientist who pores over data to ensure its accuracy can be appropriately evaluated as a meticulous researcher based on it. Being a meticulous researcher is an evaluation of who she is as an agent. It is not an evaluation of her action, nor is it an evaluation of her as a producer of actions with a certain quality, e.g., as a producer of scientifically good actions.[8]

Attributable actions of moral significance express an agent's moral identity in particular, making it appropriate to evaluate her as a moral agent. Evaluations of an agent *as a moral agent* include any moral evaluation that goes beyond appraising her merely for the moral quality of her actions. Hence, for example, they do not include appraising her as morally bad just because she produces morally bad actions.

Which appraisals go beyond appraising an agent merely for the moral quality of her actions? The most common are virtue and vice character assessments. For example, if the above willing addict takes the drug because he values it more than providing for his children, he can reasonably be evaluated as selfish or cruel for taking the drug. Yet not all appraisals of agents as moral agents are virtue and vice character appraisals. After all, an action that expresses an agent's moral identity may just express that she is minimally decent, and so not vicious or virtuous, in some respect. Further, people sometimes act "out of character," and when they do, their actions often express their moral identities.

[7] Watson (1996/2004), p. 233.
[8] My discussion in this paragraph and the next two closely follows my discussion in Section 1 of "Attributability, Weakness of Will, and the Importance of Just Having the Capacity," *Philosophical Studies*, forthcoming.

This is evident from the fact that it is often appropriate to evaluate them for such actions in virtue and vice language in a temporally bounded way. For example, if your normally unselfish friend dominates the conversation with the guest of honor, barely letting you get a word in, you may reasonably think, "that was selfish of her." In thinking this, you are thinking that she was selfish on that occasion. Her action expresses her moral identity on that occasion, even though it does not express her character.[9] Since evaluations of an agent as a moral agent are often put in virtue and vice language, I will call them "appraisals in terms of virtues and vices broadly construed." Yet keep in mind that these appraisals include any moral appraisal of an agent that goes beyond appraising her merely for the moral quality of her action.

I have just shown that an action expresses an agent's moral identity, and so is morally attributable to her, if and only if it is appropriate to appraise her for it in terms of virtues and vices broadly construed. You may wonder whether the appropriateness of these appraisals licenses further responses toward attributionally responsible agents, such as reactive attitudes or sanctions, especially since attributability is a type of responsibility.[10] On my view, like Watson's, reactive attitudes and sanctions are appropriate only toward accountable agents, and so merely being attributionally responsible for an action is insufficient for the appropriateness of these common responsibility responses.[11] Yet, for my purposes, we can safely set aside the question of whether attributability licenses further responses. First, even if it does not, the appraisals licensed by attributability are an integral part of our responsibility practice, since one way

[9] My point is that your friend's action may express her moral identity without expressing a standing disposition. Yet her action may express her character in the sense of expressing that she is not perfectly selfless.

[10] I thank an anonymous referee for pressing me on this issue.

[11] See Watson (1996/2004), pp. 230-231, 238.

to praise or blame an agent is simply to appraise her positively or negatively for her action.[12] Hence attributability is an important type of responsibility whether or not it licenses further responses. Further, to argue for my view of attributability, I just need attributability's connection to appraisals, captured by this biconditional: an agent is attributionally responsible for an action if and only if it is appropriate to appraise her for it in terms of virtues and vices broadly construed. I will use this biconditional first to argue that attributable actions are actions that are up to the agent and then to argue for the JRV.

2. Attributability and Agential Control

In this section, using the connection between attributability and the appropriateness of appraisals in terms of virtues and vices broadly construed, I make the case that attributable actions are actions that are "up to" the agent, in the sense of her having control over the fact that she performs them.[13] My argument lends support to the JRV, which cashes out this idea, and will be reinforced by my subsequent argument for the JRV.

To start, an agent's acting intentionally is necessary for attributability, since it is necessary for appropriately appraising her in terms of virtues and vices broadly construed.[14] One might object that it is often appropriate to attribute a vice to someone

[12] Watson (1996/2004) also makes this point on pp. 230-231.
[13] Because I restrict to actions of moral significance, my argument more narrowly shows that morally attributable actions are actions of moral significance that are up to the agent. But I see no reason that the connection would not hold generally.
[14] By "intentional action," I have in mind Davidson's understanding, not Velleman's. Thus, for an action to be intentional, it need only be caused by an intention. It need not, for example, result from a conscious decision. See Donald Davidson. "Actions, Reasons, and Causes," in his *Essays on Actions and Events*,

based on what she does unintentionally as a result of culpable negligence – for example, for accidentally crashing into your car while texting behind the wheel. I agree. However, in these cases, I claim that only *derivative* appraisal of an agent is appropriate for what she does unintentionally. By this, I mean that it is only appropriate to appraise her for what she does unintentionally in virtue of its being appropriate to appraise her for the intentional negligent action that led to it. This is because what the agent does unintentionally reveals no more about her moral identity than the intentional negligent action that led to it, and so it is appropriate to fundamentally appraise her only for the latter. For example, it is appropriate to appraise the driver for crashing into your car only in virtue of its being appropriate to appraise her for texting while driving, since her crashing into your car says no more about her disregard for other drivers than her texting behind the wheel does. Only the appropriateness of fundamental appraisal in terms of virtues and vices sheds light on (fundamental) attributability, and so we can safely ignore cases where only derivative appraisal is appropriate.[15]

Although acting intentionally is necessary for attributability, the above unwilling addict case shows that it is insufficient. This unwilling addict intentionally takes the drug, but he is not attributionally responsible for it, since he is a victim of his compulsive desire. We can also see this point by noting that acting intentionally is insufficient for the appropriateness of appraisal in terms of virtues and vices broadly construed. For example, although both the willing and unwilling addicts are "morally bad" in the sense of acting in a morally bad way, only the willing addict can be appropriately appraised in

---

Oxford, 2001, and J. David Velleman. "The Story of Rational Action," in his *The Possibility of Practical Reason*, Oxford, 2000.

[15] I think that agents are derivatively attributionally responsible for those actions for which it is reasonable to derivatively appraise them in terms of virtues and vices broadly construed, but I will not discuss derivative attributability here.

terms of vices for taking the drug. After all, if he takes the drug willingly because he values it more than providing for his children, he can reasonably be evaluated as selfish or cruel for taking the drug, but it seems inappropriate to evaluate the unwilling addict in terms of vices for taking the drug, as he fights helplessly against his compulsive desire.

So if attributable actions are not mere intentional actions, what are they? The answer, I claim, is that attributable actions are actions that are up to the agent, in the sense that he has control over the fact that he performs them. The above willing addict's drug use is up to him – he has control over the fact that he takes the drug – because his willingness to take the drug motivates him to take it. Because his drug use is up to him, he is attributionally responsible for it. The unwilling addict, on the other hand, is helpless in the face of his compulsive desire. He does his best to resist it, but it overcomes him, making him a victim of it. Hence he does not have control over the fact that he takes the drug. It is not up to him. Therefore, he is not attributionally responsible for it.

As the willing addict makes clear, an action can be up to an agent – he can have control over the fact that he does it – without its being the case that he could have performed a different action instead. To understand this, compare the willing addict to a drug user who is not addicted. Unlike the willing addict, the non-addict can take or not take the drug. Suppose that he takes it. He now has something in common with the willing addict: his willingness to take the drug causes him to take it. I am using "up to him" in the sense that captures this commonality between the willing addict and the non-addict, rather than in the sense of "could have done otherwise."[16]

[16] T.M. Scanlon uses "up to" in a similar sense. See T.M. Scanlon. *What We Owe to Each Other*, Belknap Press, 1999, p. 22.

Of course, another commonality between the willing addict and the non-addict is simply their willingness to take the drug. Yet being willing to take the drug while taking it is not enough to make their drug use up to them. To see this, consider a willing addict whose compulsive desire alone motivates him to take the drug. (Unlike the above willing addict, this willing addict is not motivated to take the drug by whatever represents his willingness to take it.) Although he is happy to take the drug, he does not have control over the fact that he takes it because it results only from his compulsive desire.

Notice that if attributable actions are actions that are up to the agent, as I claim, then the willing addict who is motivated by his compulsive desire alone is not attributionally responsible for taking the drug. You might worry that this is the wrong result. After all, isn't it appropriate to appraise him as selfish for taking the drug because he is willing to take it?

In my view, this willing addict is not attributionally responsible for taking the drug. Here is why: his drug use, in being caused only by a compulsive desire, does not *express* his willingness to take the drug, since his willingness to take the drug plays no role in bringing about his taking the drug. Instead, his drug use merely aligns with his willingness. The above reasoning goes wrong in thinking that it is appropriate to appraise this willing addict as selfish for taking the drug. It is appropriate to appraise him as selfish, but that is for *his willingness to take the drug* – for example, for the fact that he values taking the drug more than providing for his children – not for *taking the drug*. After all, from the fact that his compulsive desire alone causes him to take the drug, we cannot infer anything about whether he endorses taking the drug. Hence his drug use reveals nothing about his moral identity, making it inappropriate to appraise him for it.

Here I diverge from Harry Frankfurt. According to Frankfurt, an agent is (attributionally) responsible for an action that results from a desire with which he *identifies*, where an agent identifies with a desire *x* in virtue of having a higher-order desire that *x* move him to action.[17] On this view, an agent can identify with a compulsive desire and so be attributionally responsible for an action that results just from that compulsive desire. The idea is that an agent makes a compulsive desire part of his moral identity by identifying with it, and so actions that result from that compulsive desire express his moral identity. For the above reason, I think that Frankfurt is wrong on this point. Consider again the willing addict whose compulsive desire alone motivates his drug use. On Frankfurt's view, what makes him willing is that he has a higher-order desire that his desire for the drug moves him to action. However, because his drug use is caused only by a compulsive desire, this higher-order desire *plays no role* in bringing about his drug use. As such, his taking the drug does not express his higher-order desire and so does not express his willingness. It merely aligns with his willingness. The upshot is that his drug use aligns with but does not express his moral identity, and so contrary to Frankfurt, he is not attributionally responsible for it. Therefore compulsive desires are not part of an agent's moral identity, whether or not he identifies with them.

I have just argued that attributable actions are actions that are up to the agent. This means that we can think of attributability as a kind of strong agency, more robust than mere intentional agency. Hence I agree with Watson that an attributionally responsible agent is "an agent in a strong sense, an author of her conduct."[18] As I have fleshed it out, an agent in this strong sense is one who has control over the fact that she

[17] Frankfurt (1971/2003). Frankfurt's later view is more complex, but my point still holds with respect to it.
[18] Watson (1996/2004), p. 229.

performs an action. By virtue of having this control, an agent is attributionally responsible for her actions; they are attributable to her.

Because attributable actions are actions that are up to the agent, the right view of attributabiility will cash out what it takes for an action to be up to an agent. As we will see, the JRV does this. My argument for the JRV will also lend further support to the idea that attributable actions are those that are up to the agent.

3. The Judgment Responsiveness View

I propose the following view of attributability:

> **Judgment Responsiveness View (JRV):** an agent is attributionally responsible for an action A if and only if A results from either 1) his responding to at least one of his judgments about the (normative) reasons that he has in favor of doing A by doing A, *or* 2) his failing to exercise his capacity to respond to his judgments about the (normative) reasons that he has against doing A by not doing A.

Let me clarify the JRV in six ways. First, the action A may be an omission. This will be important in assessing attributability in weakness of will cases.

Second, the JRV claims that an agent makes judgments about the (normative) reasons that he has for or against performing an action. By this, I mean that an agent takes himself to have certain reasons, or sees certain reasons, for or against performing an action, and I will sometimes use these locutions instead. Importantly, an agent's judgments about normative reasons need not be correct. Thus his seeing certain reasons does not presuppose that he recognizes actual reasons or even that he has the capacity to

do so.[19] This is as it should be, since as I argued above, attributability does not require normative competence. Yet, for ease of exposition, I will sometimes say that an agent "responds to a reason that he judges/takes/sees himself to have." But we should remember that he may be wrong – a reason that he takes himself to have may only be an apparent reason; in which case, he is not responding to an actual reason but only to his taking himself to have one.

Third, an agent's judging or taking himself to have a reason should be understood functionally. If, in deciding what to do, a particular consideration plays the role of a reason – i.e., if he treats it as counting for or against an action – then he takes it to be a reason. He need not conceptualize it as a reason. To illustrate, suppose that I see my friend's favorite flowers and think, "those would cheer her up!" If, in deciding whether to buy the flowers, I take the consideration that the flowers would cheer her up as counting in favor of buying them, then I judge that this consideration is a normative reason to buy the flowers. I need not explicitly think: "that the flowers would cheer her up is a reason to buy them" or even "that the flowers would cheer her up counts in favor of buying them." In fact, I can have such thoughts without having a judgment about normative reasons, since I may explicitly think "*x* is a reason to do A" but may give *x* no weight in deciding whether to do A.

Fourth, an agent *responds* to a judgment about a reason in favor of doing some action A by doing A if and only if that judgment (sufficiently) motivates him to do A. We could equivalently say that an agent *acts on* his judgment that he has a reason in

[19] Hence the JRV differs from views of responsibility, such as that held by Fischer and Ravizza (1998), that appeal to reasons responsiveness. The JRV focuses on responsiveness to *judgments* about reasons, which may not be correct, whereas these other views focus on responsiveness to *actual* reasons. I thank an anonymous referee for pressing me to clarify this difference.

favor of doing A by doing A. Thus if I buy the flowers motivated by the consideration that the flowers would cheer up my friend, I respond to my judgment about a reason in favor of buying the flowers by buying the flowers.

Fifth, because I understand an agent's judgments about reasons functionally, these judgments may not be conscious, and the actions that result may not be either. For example, habitual actions or those done on "autopilot" result from subconscious judgments about reasons. If I am absorbed in thought while walking my usual route to work, I still take myself to have reasons to make certain turns and respond by making them, even if I do not notice this. The same is true in morally relevant cases. Suppose that I have such an entrenched habit of interrupting people that I do not consciously notice when I do it. I likely still take myself to have a reason to speak over someone, such as that I have more interesting things to say, and respond to it by interrupting. After all, if I finally notice the habit, I can ask myself why I do it and reasonably expect an answer. Hence, the JRV says – rightly in my view – that agents are attributionally responsible for subconscious actions that result from subconscious judgments about reasons. Notice that this is so whether or not the subconscious judgments about reasons are traceable to previous conscious decisions. This is because a particular consideration can subconsciously play the role of a reason in an agent's deliberation whether or not its playing that role can be traced to a previous conscious decision.

Finally, condition 2) of the JRV requires clarification. To start, it has a minor scope ambiguity. To be clear, condition 2) is: his failing to {exercise his capacity to respond to his judgments about the (normative) reasons that he has against doing A by not doing A}. Further, condition 2) is not met if an agent *lacks* the capacity to respond to

his judgments about reasons against doing A by not doing A. In order to fail to exercise that capacity, I am supposing that he must have it.

With these clarifications in hand, let me emphasize a few of the JRV's important features. To start, as we will see in Section 5, the attributability-relevant capacity proposed by the JRV – the agent's capacity to respond to his judgments about reasons – is significantly less robust than the attributability-relevant capacities proposed by other theories, which are all versions of the deep self view. In fact, we should think of the JRV as a departure from deep self views altogether. Next, on the JRV, an agent need not exercise his capacity to act on his judgments about reasons in order to be attributionally responsible for an action. This is due to condition 2), which captures the following intuition: we are attributionally responsible for actions that we could have avoided, since failing to avoid them when we could have avoided them expresses what we are like morally. As I show in Section 4, the JRV accounts for attributability for weakness of will because of condition 2), and as I discuss in Section 5, other views of attributability cannot account for attributability for weakness of will because they lack a condition like condition 2). Finally, the JRV cashes out the idea that attributable actions are those that are up to the agent. When an agent acts on his judgment about reasons, his action A is up to him; he has control over the fact that he performs A. Further, when he does A because he fails to exercise his capacity to act on his judgments about reasons against doing A by not doing A, he has the ability to avoid A, which gives him control over the fact that he performs A. So A is up to him.

Before arguing for the JRV, I should say more about the capacity in condition 2).[20] First, I simply assume that we can have but fail to exercise rational capacities. Thus I do not explain *how* we can have but fail to exercise the capacity to respond to our judgments about reasons. While others have sought to explain how we can have unexercised rational capacities, my argument does not depend upon any particular explanation of how this is so.[21] (Notice that the commonsense view of weakness of will also assumes that we can have an unexercised rational capacity – in its case, the capacity to act on our judgment about what is best.)

Second, I understand the capacity in condition 2) as a specific capacity rather than a general capacity. In other words, it is the capacity to respond to our judgments about reasons in favor of or against an action *in our specific circumstances*. It is not simply the capacity to generally respond to our judgments about reasons in favor of or against our actions or even in favor of or against some particular action. The specific capacity interpretation is the right one because, as I have argued, attributability is a kind of agential control. Being attributionally responsible for an action A requires that we have control over the fact that we perform A, and we do not have that kind of control unless we can respond to our judgments about reasons in favor of or against A in our particular circumstances. To see this, suppose that Tom can generally respond to his judgments about the reasons against taking a drug by not taking it. However, in some particular situation, he is unable to respond to those judgments and takes the drug as a result of a compulsive desire. In those circumstances, he does not have control over the fact that he

[20] I thank an anonymous referee for raising the issues that I address in the rest of this section.
[21] See, for example, Michael Smith. "Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion." in Sarah Stroud and Christine Tappolet (eds.) *Weakness of Will and Practical Irrationality* (pp. 17-38). Oxford: Oxford University Press, 2003.

takes the drug; his drug use is not up to him. Thus he is not attributionally responsible for it.

Since I adopt the specific capacity interpretation and since attributability is a type of responsibility, you might wonder whether attributability is compatible with causal determinism on the JRV. Roughly, according to causal determinism, the initial state of the world and the laws of nature fix the state of the world at all other times. You might think that, if determinism is true, then an agent who does not respond to his judgments about reasons against doing A in his particular circumstances *could not* have done so. If that is right, condition 2) would never hold, since an agent who does A would lack the capacity to respond to his judgments about reasons against doing A by not doing A.

I will not take a stand on whether attributability, according to the JRV, is compatible with determinism. In arguing for the JRV, I am arguing for a theory of attributability according to which an agent is attributionally responsible for an action if and only if either condition 1) or condition 2) obtains. It is a further question when these conditions obtain, if ever. On my view, condition 2) is sometimes met, and so either compatibilism or libertarianism holds with respect to our capacity to act on our judgments about reasons. Yet even if I am wrong – even if condition 2) never obtains because it is incompatible with determinism and determinism is true – it is still important to know, for understanding the nature of attributability, that agents are attributionally responsible for their actions when condition 2) is met. We can say something similar about condition 1), if it were threatened by determinism (which I doubt). After all, to determine whether or not attributability is compatible with determinism, we must first determine what makes agents attributionally responsible for their actions. In arguing for

the JRV, I take myself to be doing the latter. In fact, we can think of establishing the JRV as establishing a conceptual premise about the nature of attributability in an argument for the compatibility or incompatibility of attributability and determinism.

4. Argument for the JRV

I will now argue for the JRV using the biconditional established in Section 1.2: an agent is attributionally responsible for an action if and only if it is appropriate to appraise him for it in terms of virtues and vices broadly construed. (I will sometimes shorten the right-hand side to "appropriate to appraise him for it morally.") Specifically, using this biconditional, I will show that the JRV gets the right results for the right reasons in the hard cases of compulsion, weakness of will, and brainwashing. It will be evident from this discussion that the JRV gets the right results in straightforward cases.

4.1. Compulsion

4.1.1. Compulsion: The extreme unwilling addict

Recall Frankfurt's extreme unwilling addict, who fights helplessly against his addiction. Assume that he sees no reasons in favor of taking the drug but only reasons against it.[22] On this assumption, the JRV says that he is not attributionally responsible for taking the drug. Condition 1) is not met because he sees no reasons in favor of taking the drug and so did not respond to any such reason. Condition 2) is not met because, due

[22] I discuss unwilling addicts who see reasons in favor of taking the drug in Section 4.1.3.

to his addiction, he lacks the capacity to respond to the reasons that he sees against taking the drug by not taking it. This is the right result, since it is inappropriate to appraise the extreme unwilling addict morally for taking the drug.

You might object as follows. Suppose that Albert is a heroin addict who sees no reason to take heroin. Yet when he injects himself with heroin, he injects himself slowly rather than quickly to avoid overdose. Thus, he sees reasons to take heroin slowly and responds to them. He is therefore attributionally responsible for taking heroin slowly on the JRV. But how can he be attributionally responsible for taking heroin slowly but not for taking it?[23]

For the purposes of attributability, we must individuate actions in a fine-grained way, such that taking heroin and taking heroin slowly, although constituted by the same bodily movements, are actually distinct actions. To see that this is reasonable, consider another example. Suppose that I am driving my car when the brakes fail (through no fault of my own). I see that, if I continue straight ahead, I will crash into some bushes at the end of the street. I then remember that my obnoxious coworker lives near the end of the street, his house surrounded by a fence that runs just to the left of those bushes. At the last second, as I recall his latest caustic remark, I turn the wheel to the left and step on the gas, crashing through his fence. In this case, it is obvious that, although I cannot reasonably be appraised morally for crashing my car, since the brakes failed, I can reasonably be appraised morally for crashing my car through my coworker's fence. Thus I am attributionally responsible for crashing my car through my coworker's fence but not

[23] I thank Gideon Rosen for this objection.

for crashing it. So for the purposes of attributability, I can conceive of these as distinct fine-grained actions.[24]

Once we individuate actions in a fine-grained way, such that Albert's taking heroin and his taking heroin slowly are distinct actions, we see that the JRV is correct to say that Albert is attributionally responsible only for the latter. If Albert takes heroin slowly to spare his family the medical bills and humiliation that would come if he overdosed, it seems appropriate to appraise him positively for that, as it expresses concern for his family. But it is certainly inappropriate to appraise him positively for taking heroin, and it is inappropriate to appraise him negatively for it, since he is the victim of his compulsive desire. Thus, the JRV gets the extreme unwilling addict case right after all.

These examples reinforce the idea that attributable actions are those that are up to the agent. Crashing my car was not up to me, but crashing my car through my coworker's fence was. To put it in terms of control, I did not have control over the fact that I crashed my car, but I did have control over the fact that I crashed it into my coworker's fence. That is why I am attributionally responsible only for the latter, and so can be reasonably appraised only for the latter. Similarly, it is up to Albert that he takes heroin slowly but not that he takes heroin, making him attributionally responsible only for taking heroin slowly. This accounts for how it is appropriate and inappropriate to appraise him morally.

[24] Alternatively, we can say that, just as an agent can perform actions intentionally under some descriptions but not others, so he can be attributionally responsible for actions under some descriptions but not others. On this approach, I am attributionally responsible for the action under the description "crashing my car through my coworker's fence" but not for the action under the description "crashing my car."

4.1.2. Compulsion: The Willing Addict

Return to the willing addict. Above I claimed that, if his willingness motivates him to take the drug, it is appropriate to appraise him morally for it, and so he is attributionally responsible for it. The JRV yields this result. In virtue of being willing, the willing addict judges that he has reasons in favor of taking the drug. We can then reasonably understand the assumption that his willingness motivates him to take the drug to mean that he responds to a reason that he sees in favor of taking it. Thus the willing addict meets condition 1) of the JRV.

Recall that the willing addict may be overdetermined to take the drug: he may take it because he responds to his judgments about reasons in favor of taking it and because of his compulsive desire. Alternatively, he may take the drug only in response to his judgments about reasons, with the compulsive desire acting as a fail-safe device: the compulsive desire would motivate him if his judgments failed to do so. On either scenario, the willing addict responds to a reason that he sees in favor of taking the drug, and so he is attributionally responsible for taking it according to the JRV.

Yet the overdetermined willing addict may not be fully attributionally responsible for taking the drug on the JRV. That depends upon how much motivation his judgments about reasons supply – i.e., it depends upon the extent to which he responds to his judgments about reasons. If his judgments supply sufficient motivation to take the drug, he is fully attributionally responsible for taking it, just like the willing addict whose judgments alone motivate him to take the drug. However, if his judgments supply less than sufficient motivation, then he is less than fully attributionally responsible for taking the drug. His degree of attributability depends upon how much motivation his judgments

supply relative to sufficient motivation. Thus the JRV allows attributability to come in degrees.[25]

The typical willing addict's judgments about reasons provide at least some motivation to take the drug. However, what about a willing addict who is motivated only by his compulsive desire? In this case, the JRV says that he is not attributionally responsible for taking the drug. After all, he does not respond to a reason that he sees in favor of taking it, and he lacks the ability to respond to any reasons that he sees against taking it by not taking it. As I argued in Section 2, this is the correct result because, in being motivated only by his compulsive desire, his taking the drug does not *express* his willingness and so does not express his moral identity.

Yet now, using the JRV, I can say more to accommodate an intuition that this willing addict is attributionally responsible. Notice that he omits to do the following: take the drug based on the reasons that he sees in favor of taking it. (Instead, he only takes it compulsively.) The JRV correctly says that he is attributionally responsible for this omission, thus accommodating the intuition that he is attributionally responsible.

To see this, notice that we can respond to reasons in favor of or against *acting on certain reasons*. For example, this willing addict may take the drug only compulsively, rather than for the reasons that he sees in favor of taking it, because he wants his perceptive caseworker to believe that he hates taking the drug.[26] By responding to this

[25] Strictly speaking, the JRV is formulated for full attributablity. For example, condition 1) says that an agent is attributionally responsible for A if he responds to the reasons that he sees in favor of doing A *by doing A*. However, we can easily adjust the formulation to accommodate degrees of attributability. We can modify condition 1) to say: an agent is attributionally responsible for A to degree x if he responds to the reasons that he sees in favor of A *by being motivated to do A to degree x*, where x is some proportion of full motivation. We can adjust condition 2) similarly.

[26] I am not claiming that we can choose the reasons for which we act or fail to act. Rather, I am pointing out that we can have reasons not to act on certain reasons, which motivate us not to act on those reasons. I thank Angela Smith for pushing me to clarify this point.

reason in favor of the omission, the JRV says that he is attributionally responsible for it, which is correct. After all, we can reasonably appraise him as deceptive for it. To reinforce this point, notice that, although this willing addict does not have control over the fact that he takes the drug (since it results from a compulsive desire), he *does* have control over the fact that, in taking it, he does not respond to the reasons that he sees in favor of taking it.

4.1.3. Compulsion: The Typical Unwilling Addict

Unlike the extreme unwilling addict, a typical unwilling addict sees reasons in favor of taking the drug, such as that it would relieve a painful craving, and takes the drug for these reasons. If he does, the JRV says that he is attributionally responsible for taking the drug, even though we consider him to be an unwilling addict.

This case highlights that, on the JRV, the attributability-relevant capacity is very weak. An agent who acts on a single judgment about a reason that he has in favor of his action – whether he takes that reason to be strong or weak – is attributionally responsible for his action. Such a weak condition makes agents attributionally responsible for nearly all of their actions. The typical unwilling addict, who has often served as a paradigm of a non-attributionally responsible agent, seems to challenge the JRV on this point. After all, since he is an unwilling addict, it may seem inappropriate to appraise him morally for taking the drug.

The JRV gets the typical unwilling addict case right. Although harsh negative appraisal of this addict is inappropriate, it is nevertheless appropriate to appraise him morally for taking the drug. Supposing that his drug use causes serious pain to his

family, it is reasonable to say that he is less virtuous than an unwilling addict who refuses to take the drug to relieve the craving because of the pain that his drug use causes his family. Such appraisal goes beyond appraising the typical unwilling addict merely for the moral quality of his action; thus he is attributionally responsible for taking the drug, like the JRV claims.

Now consider the above-mentioned unwilling addict who refuses to take the drug to relieve the painful craving because of the pain that his drug use causes his family, and so his compulsive desire alone causes his drug use. The JRV treats this case similarly to the willing addict who takes the drug only compulsively rather than for the reasons that he sees to take it. This unwilling addict, in being motivated only by his compulsive desire, is not attributionally responsible for taking the drug, but he is attributionally responsible for the following omission: for failing to take the drug to relieve the painful craving. This is the correct result. His not taking the drug to relieve the craving expresses how much he values his family, and so he deserves positive appraisal for it. Hence he is attributionally responsible for it, as the JRV says.

4.2. Weakness of Will

A weak-willed agent has the capacity to act on his judgment that it is best to do A, but he does something else, B, instead. Is an agent attributionally responsible for his weak-willed actions? Yes. A weak-willed agent's failure to exercise his capacity to act on his best judgment expresses his moral identity. This is clear from the fact that, when we criticize an agent for being weak-willed, we are criticizing him for failing to use the self-control that he possesses, and this criticism goes beyond appraising him merely for

the moral quality of his actions. Hence weakness of will is a vice broadly construed. Because weakness of will is a vice broadly construed, agents are attributionally responsible for acting weakly.

I will not argue further for the claim that weakness of will is a vice broadly construed.[27] Instead, I will provide two examples to make it intuitively plausible.

First, imagine that a high-profile philosopher is coming to give a talk at your University, and you and your colleague are anxious to impress her. Before arriving, she sends a copy of her paper to your colleague with instructions to circulate it to those interested in attending the talk. When you next see your colleague, you ask him if he has received a copy of the paper, and he judges that it is best to tell you that he has. However, worried that the speaker would be more impressed by you than by him if you both read the paper in advance, he weakly lies, saying that he has not received it. In this case, we can reasonably criticize your colleague not just for being self-serving but also for failing to exercise the self-control that he possesses in order to tell you the truth. As this criticism is not mere criticism of him for the moral quality of his action, his weakness of will is a vice broadly construed.

Second, consider Gary Watson's example of a squash player who, after a devastating defeat, smashes his opponent in the face with his racquet even though he sees nothing valuable about doing so.[28] Whatever the attributability-relevant capacity is, assume that the squash player does not exercise it. For example, imagine that his desire is purely appetitive (and so is not the result of any evaluative stance, commitment, care, etc.), and imagine that when he acts upon that desire, he only exercises his capacity for

[27] For a more in-depth argument, see Strabbing (forthcoming).
[28] Gary Watson. "Free Agency," in his (ed.) *Free Will*, Oxford University Press, 2003, p.19. (Original work published 1975)

intentional action (which we have seen is insufficient for attributability). Now assume that the squash player is not compelled to act on this desire: he could have exercised his attributability-relevant capacity to avoid smashing his opponent in the face. Because of this, it is appropriate to appraise him as hot-headed for smashing his opponent in the face, which is to say that he failed to use self-control in a fit of temper. Since this criticism is not mere criticism of the moral quality of his action, his hot-headedness – his weakness of will – is a vice broadly construed.

Does the JRV say that agents are attributionally responsible for their weak-willed actions? It does, thanks to condition 2).[29] Consider Watson's squash player, who illustrates how condition 2) alone can be met. Although a typical squash player who smashes his opponent in the face sees some reasons to do so and acts on them, we are imagining the rare case in which the squash player does not exercise his capacity to act on his judgments about reasons. Instead, he acts on a desire to smash his opponent in the face that is disconnected from his judgments about reasons. It could be either that he sees no reason to smash his opponent in the face or that the reasons that he sees to smash his opponent in the face are motivationally inert. Either way, because he acts on that disconnected desire, condition 1) is not met. Yet, via condition 2), the JRV says that this squash player is attributionally responsible for smashing his opponent in the face, since he could have responded to the reasons that he sees against smashing his opponent in the face by not smashing his opponent in the face. This is the right result because his acting on that resistible desire shows weakness of will.

[29] The argument from this point through the end of the section is a specific case of the more general argument that I make in Section 5 of Strabbing (forthcoming) for the Having the Capacity Principle, a general principle of which the JRV is a specific instantiation. See footnote 37 below for a brief discussion of the Having the Capacity Principle and how the JRV is an instantiation of it.

Yet, importantly, condition 2) of the JRV does not just come into play in rare cases like the squash player; it is crucial for accounting for attributability for weakness of will as such. To see this, consider the JRV specifically for omissions:

> **JRV (omissions):** an agent is attributionally responsible for not doing A if and only if his not doing A results from either 1) his responding to at least one of his judgments about the reasons that he has in favor of not doing A by not doing A, or 2) his failing to exercise his capacity to respond to his judgments about the reasons that he has in favor of doing A by doing A.

Now notice that weakness of will itself is an omission: the agent's not acting on his best judgment (when he could have acted on it). Condition 2) says that agents are always attributionally responsible for this omission. After all, if an agent judges that some action is best, then he judges that he has reasons in favor of performing that action. Hence, in acting weakly, his not performing the action that he judges best results from his failure to exercise his capacity to respond to his judgments about the reasons that he has in favor of performing the action that he judges best by performing that action. Thus condition 2), but not condition 1), of the JRV says that agents are always attributionally responsible for weakly failing to do the action that they judge best. This is the right result because this omission is precisely what makes it appropriate to appraise an agent as weak-willed.

Of course, in specific cases, a weak-willed agent judges that it is best to do A but does B instead, and so his not performing the action that he judges best is constituted by his failure to do A, which in turn is constituted by his doing B. Condition 2) of the JRV says that weak-willed agents are attributionally responsible for their failure to do A. For example, condition 2) says that your colleague is attributionally responsible for weakly failing to tell you the truth about his receiving the paper, since his failure to tell you the

truth results from his failure to exercise his capacity to respond to the reasons that he sees in favor of telling you the truth by telling you the truth.

Things are more complicated with respect to the weak-willed action B. In performing B, weak-willed agents often respond to their judgments about reasons in favor of doing B by doing B, thus meeting condition 1) of the JRV. For example, in lying to you about receiving the paper, your colleague responds to a reason that he sees in favor of lying to you: that the senior philosopher will be more impressed by you than by him if he gives you the paper. However, it is important to recognize that condition 2) *also* says that agents are attributionally responsible for the weak-willed action B. For example, your colleague is also attributionally responsible for lying to you because it results from his failure to exercise his capacity to respond to the reasons that he sees against lying to you by not lying to you. Further, B is a weak-willed action precisely because it constitutes the agent's failure to act on his best judgment (when he could have acted on it), and so on the JRV, an agent is attributionally responsible for B *as a weak-willed action* because condition 2) is met with respect to it. I think that the JRV gets the right results here. We can appropriately appraise your colleague as self-serving for lying because of the reasons that he responds to in lying, which is captured by condition 1), but we can appropriately appraise him as weak-willed for lying because condition 2) is met.

4.3. Brainwashing

Consider *brainwashing* (as philosophers typically imagine it). Brainwashing is distinct from hypnosis: while under hypnosis, you cannot respond to the reasons that you see for or against your actions because the hypnotist's instructions bypass your rational

faculties.[30] So, on the JRV, you are not attributionally responsible for actions performed under hypnosis, and this seems right. However, unlike hypnosis, which disconnects you from your judgments about reasons, brainwashing *changes* your judgments about reasons. On the JRV, you are attributionally responsible for actions that result from these inculcated judgments about reasons.

Although this may initially seem implausible, it is the right result. Brainwashing changes the agent's moral identity, and the resulting actions express his moral identity after the brainwashing. His actions are up to who he now is. For this reason, he can legitimately be morally appraised for them. For example, if someone steals a car because he is brainwashed into thinking that wanting something is a good reason to steal it, we can legitimately think that he has become egotistical and malicious from the brainwashing. This shows that attributability does not depend upon *how* a person got to be the way he is morally but simply on what he is like morally. (How a person got to be the way he is morally seems relevant to another type of responsibility: accountability.)[31]

5. We Should Reject Deep Self Views of Attributability

So far I have argued that attributable actions are those that are up to the agent and that the JRV, which cashes out this idea, is the right view of attributability. In this section, I demonstrate that the JRV is a significant departure from the other views of

[30] If I am wrong about how hypnosis works, we can still use the phenomenon that I describe as a contrast to brainwashing.

[31] One might object that a neuroscientist could implant in someone's brain the judgment that he has some reason to steal the car, which results in his stealing it. It then seems inappropriate to appraise him morally for stealing it, even though he responds to a judgment about reasons in favor of it. In this case, however, he does not respond to *his* judgment about reasons. To be his judgment, he must form it using his rational capacity, which the neuroscientist bypasses. If the neuroscientist instead manipulates this capacity, it is brainwashing.

attributability for actions on offer.[32] First, unlike those views, the JRV is not reasonably thought of as a deep self view, and so attributable actions should not be conceptualized as actions expressive of the deep self. Second, unlike those views, the JRV – via condition 2) – says that we are attributionally responsible for actions that result from failing to exercise the attributability-relevant capacity to avoid them. As a result, only the JRV adequately accounts for weakness of will.

Regarding the first difference, consider again the typical unwilling addict who takes the drug to relieve the painful craving. I have argued that he is attributionally responsible for taking the drug, as the JRV says. Yet deep self views of attributability get this case wrong. On deep self views, the typical unwilling addict is not attributionally responsible for taking the drug because it does not express "deep" features such as his fundamental values, cares, or commitments. As Watson says, for deep self views, "what is in question is an individual's fundamental evaluative orientation,"[33] but taking the drug does not express the typical unwilling addict's fundamental evaluative orientation. Because all views of attributability for actions on offer are deep self views, unwilling addicts have been paradigms of agents who are not attributionally responsible for their actions. The JRV shows that they should no longer serve as such paradigms. Only the extreme unwilling addict is not attributionally responsible for taking the drug.

---

[32] Angela Smith defends a somewhat similar account of attributability (the "rational relations view"), but her primary concern is attributability for attitudes rather than for actions, and she defends her view in the context of arguing against volitional views of responsibility. Further, her view does not have a condition analogous to condition 2). Finally, she does not understand attributability as the "up to" relation, since she thinks that we are directly attributionally responsible for mental acts like forgetting in virtue of the fact that they reflect objectionable attitudes. See Angela M. Smith. "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138 (3):367-392, 2008; Angela M. Smith. "Responsibility for Attitudes: Activity and Passivity in Mental Life," *Ethics* 115 (2): 236-271, 2005; Angela M. Smith. "Conflicting Attitudes, Moral Agency, and Conceptions of the Self," *Philosophical Topics* 32 (1/2):331-352, 2004.

[33] Watson (1996/2004), p. 234.

Deep self views get the typical unwilling addict case wrong because they construe an agent's moral identity too narrowly. An agent's fundamental values, cares, and commitments are perhaps the most important components of an agent's moral identity, but they do not fully constitute it. Otherwise, the moral identity of a perfectly virtuous person would be indistinguishable from the moral identity of someone with the right fundamental values, cares, and commitments but who nevertheless has and acts on occasional petty jealousies and selfish concerns. Since the moral identities of these agents are distinguishable, we need a broader conception of an agent's moral identity than that provided by deep self views. We see how broad that conception should be once we link attributability to appraisal in terms of virtues and vices broadly construed, which allows us to make the required fine discriminations amongst moral identities.

You might object that the JRV is also a deep self view on the grounds that an agent's deep self just is his moral identity. Since attributable actions express an agent's moral identity, any theory of attributability would then be a version of the deep self view. The question would be whether a particular theory of attributability successfully captures the contours of the deep self.

We could identify the agent's deep self with his moral identity, making the JRV a deep self view. But importantly, the JRV would still represent a significant departure from the way in which the deep self has been understood. Specifically, the JRV would show that the boundary of the deep self is much less "deep" than has previously been thought, extending well beyond an agent's fundamental values, cares, or commitments to include all of the agent's judgments about reasons, no matter how fleeting the judgment or how opposed to his deep values, cares, or commitments.

Yet, to my mind, thinking of the JRV as a deep self view stretches the notion of the deep self too far. If you act on a fleeting judgment about a reason in favor of an action, a judgment opposed to your fundamental values, it sounds wrong to say that this action expresses your deep self. This is reinforced by Gary Watson's claim that, on deep self views, attributable actions are actions that express an agent's adopted ends.[34] It is implausible that an agent adopts an end whenever he judges that there is a reason in favor of or against an action. After all, agents often see reasons both for and against particular actions, and it is implausible that, in weighing the pros and cons of some action, agents adopt as ends all of the considerations that they take to be reason-giving. Finally, lumping all of our judgments about reasons into the deep self obscures an important point: that our moral identities are composed of features on a spectrum from deep to shallow. As I mentioned above, having the same deep values, cares, and commitments as a virtuous person does not make you a virtuous person. Contrasting the JRV with deep self views highlights the fact that our "shallow" cares, concerns, and judgments also partly constitute our moral identities.

Interestingly, contrasting the JRV with deep self views draws attention to the important role played by an agent's fundamental values, cares, and commitments. An attributable action that expresses the deep self expresses something more significant about the agent's moral identity than an attributable action that expresses the "shallow self." Contrast the above-mentioned typically unselfish friend who dominates the conversation with the guest of honor and a selfish friend who does the same. Only the latter's action expresses her deep self, and so it expresses something more significant about her – that she is a selfish friend – than the typically unselfish friend's action, which

[34]Watson (1996/2004), p. 228.

expresses that she is not as selfless as she could be. Hence an agent is more blameworthy in the attributability sense for a bad action – i.e., the action expresses something worse about him – the more the action expresses his deep self.

Further, when an agent performs an action that expresses his deep self, he will likely be attributionally responsible for more fine-grained actions than an agent whose action does not express his deep self.[35] Consider again the willing addict and the typical unwilling addict. The willing addict may take the drug for multiple reasons – e.g., he enjoys it, he thinks that it makes him look cool, and he does not see himself as responsible for his family. Thus he is attributionally responsible not just for the course-grained action of taking the drug but also for the fine-grained actions of taking it for enjoyment, of taking it to look cool, and of taking it because he does not see himself as responsible for his family. The typical unwilling addict who only takes the drug to relieve the painful craving is not attributionally responsible for so many fine-grained actions. Hence, if a bad action expresses the deep self, the agent will likely be more blameworthy in the attributability sense in that he will likely be blameworthy for more fine-grained actions.[36]

For the above reasons, I think that we should understand the JRV as a departure from deep self views. Contrary to deep self views, attributability is not a relationship between an agent's action and his fundamental values, cares, or commitments. Instead, it is a relationship between an agent's action and his judgments about reasons.

[35] Alternatively, as we saw above, we can say that when an agent performs an action that expresses his deep self, he will likely be attributionally responsible for it under more descriptions than for an action that does not express his deep self.

[36] I thank Reed Winegar for helpful discussion on this point.

The JRV is also a significant departure from the other views of attributability in incorporating condition 2). Condition 2) is a version of the following sufficient condition for attributability: an agent is attributionally responsible for an action A if A results from his failure to exercise the attributability-relevant capacity to avoid doing A. The idea is that we are attributionally responsible for actions that we could have avoided, since our not avoiding them when we could have expresses what we are like morally. I argue elsewhere for this sufficient condition and show that current views of attributability fail to recognize it, thus failing to account for attributability for weakness of will. (Instead, these views take *exercising* the attributability-relevant capacity to be necessary and sufficient for attributability.)[37] Although I cannot repeat that argument here, we can see the point just from the fact that the JRV needs condition 2) to account for attributability for weakness of will. As we saw above, only condition 2) says that agents are attributionally responsible for their failure to perform the action that they judge best, the defining omission of weakness of will. Further, when an agent judges that it is best to do A but does B instead, only condition 2) says that he is attributionally responsible for not doing A and that he is attributionally responsible for B *as a weak-willed action*. Without

[37] See (Strabbing, forthcoming). There I argue that being attributionally responsible for an action is not a matter of exercising the attributability-relevant capacity but is simply a matter of having it (so long as having that capacity figures in the explanation of the action). Specifically, I argue that the correct view of attributability must embrace the following general principle:

> Having the Capacity Principle (HC Principle): an agent is attributionally responsible for an action A if and only if 1) A results from the exercise of his attributability-relevant capacity to do A or 2) A results from his failure to exercise his attributability-relevant capacity to avoid doing A.

As I show in that paper, current views of attributability do not accept the HC Principle because they do not accept its condition 2), and so they cannot account for attributability for weakness of will. The JRV avoids this mistake. Notice that the JRV is a specific instantiation of the HC Principle, in which the attributability-relevant capacity is the agent's capacity to respond to his judgments about reasons. Ultimately, the JRV accommodates attributability for weakness of will because it embraces the HC Principle – specifically, condition 2) of the HC Principle.

an analogue to condition 2), and so without embracing the above sufficient condition, the other views fail to give these results and so fail to account for attributability for weakness of will. Of course, these views could incorporate the above principle and rectify the problem. Yet currently, condition 2) represents a second way in which the JRV is superior to all other views of attributability.

6. Conclusion

I have argued for the following view of attributability for actions:

> **Judgment Responsiveness View (JRV):** an agent is attributionally responsible for an action A if and only if A results from either 1) his responding to at least one of his judgments about the reasons that he has in favor of doing A by doing A, *or* 2) his failing to exercise his capacity to respond to his judgments about the reasons that he has against doing A by not doing A.

As I demonstrated, the JRV yields the right answers in the hard cases of compulsion, weakness of will, and brainwashing, saying that an agent is attributionally responsible for his action when and only when he can be reasonably appraised for it in terms of virtues and vices broadly construed. Further, the JRV cashes out the idea that attributable actions are actions that are up to the agent, an idea for which I argued independently. Thus my argument for that idea supports my argument for the JRV and vice versa.

As I have argued, the JRV is a significant departure from the other views of attributability in two respects. First, it is not reasonably described as a deep self view. In claiming that an agent's moral identity is constituted by his judgments about reasons, only the JRV recognizes that an agent can be attributionally responsible for an action

even if it does not result from his fundamental values, cares, or commitments because our moral identities are broader than these elements. Thus we should not conceptualize attributable actions as actions expressive of the deep self. Second, unlike other views, the JRV recognizes that we are attributionally responsible not just for actions that result from exercising the attributability-relevant capacity but also for actions that result from *failing* to exercise the attributability-relevant capacity to avoid them. As a result, only the JRV can account for attributability for weakness of will. The two ways in which the JRV diverges from the other views of attributability show that these other views are too narrow. Attributability is a broader and richer conception of responsibility than previously thought.

Acknowledgments

I thank Lara Buchak, D. Justin Coates, Gideon Rosen, Angela Smith, Michael Smith, and Gary Watson for very helpful comments and discussions on previous drafts of this paper. I also thank the participants of the University of San Francisco Conference on Responsibility, Agency, and Persons 2 for valuable discussion on a previous draft of this paper.