Jada Twedt Strabbing
Fordham University
Penultimate draft. Paper forthcoming in *Philosophical Studies*

Attributability, Weakness of Will, and the Importance of Just Having the Capacity

Abstract:

A common objection to particular views of attributability is that they fail to account for weakness of will. In this paper, I show that the problem of weakness of will is much deeper than has been recognized, extending to all views of attributability on offer because of the general form that these views take. The fundamental problem is this: current views claim that being attributionally responsible is a matter of exercising whatever capacity that they take to be relevant to attributability; however, weakness of will cases show that we can be attributionally responsible for actions that result from failing to exercise that capacity. I propose a novel solution that any view of attributability can and must take on board in order to be viable. The solution is to recognize that being attributionally responsible is not fundamentally a matter of exercising the attributability-relevant capacity (whatever a particular view identifies that capacity to be) but is rather a matter of *having* that capacity, so long as that capacity figures in the explanation of the action.

A common objection to particular views of attributability is that they fail to account for weakness of will. In this paper, I show that the problem of weakness of will is much deeper than has been recognized, extending to all views of attributability on offer because of the general form that these views take. I then propose a novel solution that any view of attributability can and must adopt in order to be viable.

Following Gary Watson (1996/2004), I take attributable actions to be those actions that express what an agent is like practically – i.e., that express his practical identity. A particular view of attributability fleshes out what it takes for an action to express an agent's practical identity. To be viable, it must account for weakness of will. An agent is weak-willed if and only if he fails to act on his judgment about what is best

despite having the capacity to act on that judgment.[1] It is reasonable to think that weak-willed agents act freely in the sense required for attributability and hence can be attributionally responsible for their weak-willed actions. Particular views of attributability have been challenged on the grounds that they do not give this result.

Yet the fundamental problem – the problem that weakness of will poses for all current views of attributability – has not been recognized. The fundamental problem is this: current views claim that being attributionally responsible is a matter of exercising the capacity that they take to be relevant to attributability; however, weakness of will cases show that we can be attributionally responsible for actions that result from *failing* to exercise that capacity. Fortunately, the solution to this problem is one that any view of attributability can adopt. The solution is to recognize that being attributionally responsible is not fundamentally a matter of exercising the attributability-relevant capacity (whatever a particular view identifies that capacity to be) but is rather a matter of *having* that capacity.

My argument proceeds as follows. In Section 1, I show that an agent is attributionally responsible for an action (of moral significance) if and only if it is appropriate to appraise him for it in terms of virtues and vices broadly construed. This biconditional will allow us to identify when an agent is and is not attributionally responsible for an action. Then, in Section 2, I move toward the fundamental problem of weakness of will by examining John Martin Fischer's recent criticism of "value-added" views of attributability (Fischer 2010). In particular, I consider his criticism applied to a specific value-added view, namely Gary Watson's (1975/2004), and I demonstrate how Watson can broaden his view to overcome it. However, I show in Section 3 that the broadened view still cannot account for weakness of will and, in Section 4, that this problem generalizes to all views of attributability on offer. Finally, in Section 5, I present and argue for my solution to the generalized problem.

[1] I set aside the issue of how to flesh out the capacity to act on our best judgment such that we can have it but fail to exercise it. For my purposes, what matters is that we have such a capacity – which we intuitively do – not that it has a particular form. (Others – for example, Smith (2003) – have offered views about how to understand the ability to act on our best judgment such that we can have it but fail to exercise it.) Although I will not argue that this intuitive understanding of weakness of will is viable, I respond to a challenge to it below.

1. Attributability and the Connection to Virtues and Vices

As Gary Watson argues, attributable actions open the agent up to distinctive forms of evaluation – namely, evaluations of him as an agent – precisely because they express his practical identity (Watson 1996/2004, p. 233). For example, if making detailed financial spreadsheets is attributable to Lara and so expresses what she is like practically, she can be appropriately evaluated as a careful planner on account of it. Being a careful planner is an evaluation of who she is as an agent. It is not an evaluation of her action, nor an evaluation of her as a producer of actions with a certain quality – e.g., as a good producer of detailed financial data.

For actions of moral significance, attributable actions specifically express the agent's moral identity and so open him up to distinctive forms of moral evaluation – namely, evaluations of him as a moral agent. Evaluations of an agent *as a moral agent* include any moral evaluation of him that goes beyond appraising him merely for the moral quality of his actions. So, for example, they do not include appraising him as morally bad just in virtue of his producing morally bad actions.

Which moral appraisals go beyond appraising the agent merely for the moral quality of his actions? The standard ones are virtue and vice character assessments. For example, if pilfering cash from the register is attributable to Christina, it expresses her moral identity, and if she does it to buy a stylish wardrobe, it is appropriate to evaluate her as selfish or greedy for it. Yet not all appraisals that implicate an agent's moral identity are virtue and vice character appraisals. An action that expresses an agent's moral identity may just express that he is morally average – not particularly vicious or virtuous – in a certain respect. Further, a person may act "out of character," and moral appraisal of him in virtue and vice language may be appropriate in a temporally bounded way. For example, a usually selfless husband may reasonably say to his wife: "It was selfish of me to watch TV while you cleaned all afternoon to prepare for our guests. I'm really sorry." In saying this, he admits that he was selfish for acting as he did, i.e., that he was selfish on that occasion. His action expresses his moral identity, at least on that

occasion, even though it does not express his character.[2] Because appraisals of a moral agent are typically couched in virtue and vice language, even when they do not reflect standing dispositions, I will call them "appraisals in terms of virtues and vices broadly construed." It should be remembered though that these appraisals include any moral appraisal that goes beyond appraising someone merely for the moral quality of his action.

I have just shown that an action expresses an agent's moral identity, and so is morally attributable to him, if and only if it is appropriate to appraise him for it in terms of virtues and vices broadly construed. Because I am focusing on moral attributability, I can therefore use the following biconditional to discern whether an agent is attributionally responsible for action: 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. (From now on, I will often use "moral appraisal" as short for "appraisal in terms of virtues and vices broadly construed".) With this resource, let's examine Fischer's criticism.

2. Fischer's Criticism of Value-Added Views of Attributability

As Fischer defines it, "value-added" views of attributability are views that incorporate a particular type of normative or value condition into their analysis. Specifically, they claim that, to act freely in the sense required for attributability, an agent must "act in accordance with what he takes to be rationally or normatively defensible." (Fischer 2010, p. 314).[3] (By focusing on acting freely, we set aside the issue of whether other conditions, such as epistemic conditions, must also be met for an agent to be attributionally responsible for an action.) The problem for value-added views, according to Fischer, is that they cannot account for the fact that we can be attributionally responsible for weakly acting against what we take to be normatively defensible.

[2] Of course, the husband's action expresses his character in the sense that it expresses that he is not perfectly selfless. My point is that an action can express an agent's moral identity without expressing a standing disposition.

[3] Fischer actually says "value-added views of responsibility," but attributability is clearly the type of responsibility that he has in mind, since he discusses views of attributability and claims that value-added views are views of autonomy.

To understand this objection, and to see how it can be overcome, focus on Watson's original view of attributability. On this view, an agent acts freely in the sense required for attributability if and only if she acts on what she most values, where what she most values is what she judges to be the most valuable.[4] Watson's view fits Fischer's definition of a value-added view. It claims that acting freely is acting on what the agent takes to be normatively defensible, where what the agent takes to be normatively defensible is what she judges to be the most valuable in her circumstances.

Watson's view is problematic for the reason that Fischer states: it says that we are not attributionally responsible for weakly acting against what we most value. This is the wrong result. Consider a company CEO who judges that it is best to level with his shareholders about the company's poor profits but who also values his lavish lifestyle, a lifestyle which will be impossible to maintain if he is honest with his shareholders and the company's stock prices inevitably drop. Suppose now that the CEO has the capacity to resist acting on his desire to maintain his lavish lifestyle; yet, he weakly gives in and lies to his shareholders. According to Watson's view, the CEO is not attributionally responsible for lying to his shareholders. But he clearly is. It is appropriate for the shareholders who later discover the deception to evaluate him as dishonest and greedy.[5]

Watson can reasonably broaden his view to overcome this objection. Rather than claiming that an agent acts freely (in the sense required for attributability) if and only if he acts on what he *most values* in the situation, he can claim that an agent acts freely if and only if he acts on what he *values* in the situation. Thus, even if the CEO most values telling the truth, he is still attributionally responsible for lying to his shareholders on this broader view because, in doing so, he acts in pursuit of something that he values – namely, his lavish lifestyle.

---

[4] Watson (1975/2004) clearly endorses the sufficiency condition. Although less obvious, he also accepts the necessity condition. For example, he says that a free agent's actions "flow from his evaluational system." (p. 26). Further, in (Watson 1987/2004), Watson rejects his original view because "… I might fully 'embrace' a course of action I do not judge best…" (p. 168). This reason could only count against his original view if it incorporates the necessity condition. Along different lines, as I interpret Watson's original view, an agent's action must issue from, not merely align with, what he most values in order for it to be free in the sense required for attributability. But if I am wrong about this, it does not affect the points in this paper.

[5] Haji (2002) and Vihvelin (1994) also make this criticism of Watson's original view, using different examples.

This broader view is not technically a value-added view as Fischer defines it. As I interpret Fischer, weak-willed agents act against what they take to be normatively defensible. Thus the weak-willed CEO acts against what he takes to be normatively defensible but yet, on the broader view, is attributionally responsible for doing so, which is not the case for value-added views as Fischer defines them. As a result, the broader view does not fall prey to Fischer's objection: it correctly says that agents like the CEO are attributionally responsible for weakly acting against what they take to be normatively defensible – so long as they act in pursuit of something that they value. Further, this broader view is a value-added view in the clear sense of incorporating a value component. So not only can Watson overcome Fischer's objection, he can do so while reasonably claiming that his view is a value-added view of attributability.

3. The Problem with the Broader Version of Watson's View

3.1 The Problem

Unfortunately for Watson, he cannot solve the weakness of will problem so easily. The broader version of his view, which says that an agent acts freely in the sense required for attributability if and only if she acts on what she values, cannot account for attributability in all cases of weakness of will.

To see this, distinguish what I will call Type 1 and Type 2 cases of weakness of will. A weak-willed agent displays Type 1 weakness of will when (and only when) he acts weakly in pursuit of an end that he judges to be valuable. The CEO example is a Type 1 case. A weak-willed agent displays Type 2 weakness of will when (and only when) she acts weakly in pursuit of an end that she does not value at all.

To see that Type 2 cases exist, first consider that agents can desire things that they do not value. Gary Watson has provided vivid examples: the mother with the sudden urge to drown her bawling baby in the bathtub and the squash player who after suffering a crushing defeat desires to smash his opponent in the face with his racquet. As Watson

says, "[i]t is just false that the mother values her child's being drowned or that the player values the injury and suffering of his opponent. But they desire these things nonetheless" (Watson 1975/2004, p. 19). Moreover, we can surely act on such desires. Any desire that we have can motivate us to action, not just those whose objects we judge valuable. Finally, we can *weakly* act on such desires – in other words, we can act on such desires despite having the capacity to resist. As any desire can move us to action, it is reasonable to think that we can weakly act on any desire as long as weakness of will is possible, which I assume that it is. Thus, we can coherently suppose that the squash player weakly acts on his desire to smash his opponent in the face, even though he does not value doing so. This squash player, then, exhibits Type 2 weakness of will.

We have seen that the broader version of Watson's view can account for Type 1 weakness of will, like the CEO exhibits. However, it cannot account for Type 2 cases. According to the broader view, the weak-willed squash player is not attributionally responsible for smashing his opponent in the face because he does not value doing so. This is the wrong result. Agents are attributionally responsible for all weak-willed actions, including Type 2 ones. This is because a weak-willed agent has the capacity to act on his judgment about what is best, and the fact that he fails to do so despite having that capacity expresses his moral identity. I think that this is our commonsense view. When a person acts badly from weakness of will, we criticize him, possibly amongst other things, for failing to use the self-control that he possesses. Because this criticism goes beyond criticizing him merely for the moral quality of his action, it shows that failing to use self-control, or weakness of will, is a vice broadly construed. To put it another way, weak-willed actions show that the agent lacks the corresponding virtue of moral fortitude. For example, we can legitimately criticize the CEO not just for being dishonest and greedy but also for lacking the moral fortitude to act on his judgment that it is best to level with his shareholders. And the squash player, who smashes his opponent in the face even though he does not value doing so and has the capacity to resist, is rightly criticized as hot-headed – a way of saying that he failed to use self-control in a fit of temper. Because weakness of will is a vice broadly construed, weak-willed agents are always attributionally responsible for acting weakly.

The literature so far has failed to recognize that we are attributionally responsible for Type 2 weak-willed actions. Counterexamples to views of attributability utilizing weakness of will all utilize Type 1 cases, such as an akratic pie-eater who, it is reasonable to think, values the taste of the pie (Haji 2002) and a person who akratically lies to Customs officers to avoid paying import duties because, it is reasonable to think, he values keeping the money (Vihvelin 1994). Further, and more significantly, the consensus is that an agent is a passive bystander with respect to actions that result from unendorsed motives and thus that these actions cannot be free in the sense required for attributability. For example, Benjamin Mitchell-Yellin claims that "… [an agent's behavior] might issue from a motive, the object of which he does not value. This would be a case of mere behavior, to which the agent is a mere bystander" (Mitchell-Yellin 2014, p. 347).

Mitchell-Yellin and others are wrong. An agent who *weakly* acts on a desire for an object that he does not value is not a mere bystander. He does not exhibit "mere behavior." Consider the following analogy. Suppose that a man is drowning and that two bystanders on the beach notice him but do not attempt a rescue. The first bystander is a strong swimmer who has the ability to rescue the man; the second cannot swim. Both bystanders are passive in a sense: neither acts to rescue the man. However, in an important sense, only the second bystander is passive: only he lacks the ability to rescue the drowning man. This latter sense is the one relevant to attributability. After all, it is appropriate to appraise the first bystander negatively for failing to attempt the rescue, but not the second bystander. Now consider an agent who weakly acts on a desire for an end that he does not value and an agent who acts on a compulsive desire for an end that he does not value – cases that the literature fails to distinguish. With respect to attributability, these cases are crucially different. Although both are passive bystanders to their behavior in the sense that they both act on an unendorsed desire, the weak-willed agent, unlike the compulsive one, is not a passive bystander in the sense that he could have acted on his values, and this latter sense is the one relevant to attributability. After all, it is appropriate to appraise the weak-willed agent negatively for his failure to use self-control. Unlike the compulsive agent, he does not exhibit "mere behavior."

In the next section, we will see that the literature's failure to recognize that we are attributionally responsible for Type 2 cases of weakness of will is a symptom of a larger problem: its failure to see that we are attributionally responsible for weak-willed actions that do not result from exercising the capacity relevant to attributability. Before turning to this issue, I will deal with two objections to my argument so far.

3.2 Objection: Isn't Weakness of Will Morally Neutral?

One might object that, contrary to my claim, weakness of will does not express an agent's moral identity because it is morally neutral. In support, the objector can point to cases in which it seems inappropriate to appraise someone morally for a weak-willed action and cases in which positive appraisal for a weak-willed action seems appropriate. If correct, this objection would undermine my argument that weakness of will is a vice broadly construed and so undermine my argument that we are attributionally responsible for Type 2 cases.

Start with a case in which moral appraisal for a weak-willed action may seem inappropriate. Suppose that Jones desires to wear a green tie, even though he sees nothing valuable in doing so. Moreover, he falsely believes that it is morally wrong to wear green ties, but he weakly wears one. In this case, so the argument goes, it is wrong to appraise Jones negatively, since it is permissible to wear green ties, and positive appraisal is inappropriate. Thus it is inappropriate to appraise him morally at all.[6]

I disagree with this conclusion. Jones manifests a vice when he gives in to temptation and wears the green tie when he could have resisted, since his conception of his situation alone can make moral appraisal of him appropriate. To see this, imagine that Jones believes that wearing green ties to work will cause his co-workers to suffer terribly, and he wears a green tie for that reason. In this case, it is appropriate to evaluate him as cruel for wearing the tie. Now imagine instead that Jones wears a green tie precisely because he thinks that it is wrong. Doing something because it is wrong reveals viciousness, and Jones' action expresses this viciousness even though wearing green ties

[6] Thanks to *** and *** for raising this concern. The Jones example is due to ***.

is permissible. In these cases, it is appropriate to evaluate Jones morally for wearing a green tie, just as it would be if his conception of his situation were correct – if wearing green ties really were wrong. Similarly, it is reasonable to criticize Jones for failing to use self-control in the original case, since it would be reasonable to do so if wearing green ties were wrong, as he thinks.

This point is perhaps better illustrated with a less farfetched example. Suppose that Jains think that killing a fly is morally wrong, but I know that it is permissible. Now suppose that an annoying fly has been buzzing around a practicing Jain for the last hour. First, imagine that he staunchly refuses to kill it, even though it is clearly a nuisance to him. In this case, I can reasonably admire his moral fortitude, even though his belief that killing the fly is wrong is false. Now imagine instead that he swats the fly dead, even though he judged at the time that it was wrong and could have resisted. In this case, I think that I can reasonably appraise him negatively for not using his self-control to do what he thinks is right.

Turn to a case in which positive appraisal for a weak-willed action seems appropriate: Huck Finn. Huck judges that he should turn in a runaway slave, Jim, but he does not because of the affection and respect that he develops for Jim. In acting weakly, Huck does the right thing for the right reason, and so he deserves positive moral appraisal.[7] It may then seem that weakness of will is not a vice, but that does not follow. First, assuming it is a vice, our overall assessment of Huck should still be positive, since the virtue that he displays in not turning in Jim out of respect for his humanity greatly outweighs the vice of weakness of will – especially given the difficulty of acting for that reason in his slave-owning society. Hence we can accept both that Huck deserves positive appraisal and that he demonstrates a vice in acting weakly. Further, if Huck judges that he should turn Jim in,[8] it seems to me appropriate to appraise him negatively for his lack of moral fortitude. He chooses to do what he thinks is wrong, which is a negative quality. After all, it might reasonably lead us to wonder whether, in the future, he will have the moral fortitude to do what he correctly believes to be right. Of course,

[7] See Arpaly (2003) for a discussion of this case.

[8] I think that Huck likely no longer judges that he should turn Jim in. As Arpaly (2003, p. 10) describes him, Huck does not conceptualize his realization that Jim is a human being worthy of respect. Given this, it is unlikely that he would realize that his judgment about what he should do has changed. If it has changed, he does not act weakly.

given his failure to understand his situation, Huck can only do the right thing for the right reason by acting weakly. But this does not mean that his weakness of will itself is either good or neutral. If I can only do the right thing for the right reason by causing someone pain, this does not mean that my causing him pain is either good or neutral. In Huck's case, his ignorance puts him in a situation in which, on either option, he displays a vice.

With this objection dispatched, I contend that weakness of will is a vice broadly construed. It expresses something negative about the agent's moral identity: a lack of moral fortitude.

3.3 Objection: Is Weakness of Will a Species of Compulsion?

Here is a second objection, which Watson would raise. I understand weakness of will as the agent's failure to exercise his capacity to act on his best judgment. According to Watson (1977/2004), this commonsense view is wrong. Instead, he thinks that weakness of will is a species of compulsion: an agent acts weakly when he could not have resisted the desire that he acts on but a person with normal levels of self-control could have. If Watson is correct, "weak-willed" agents who act on ends that they do not value would be compelled to do so, and so it would be inappropriate to appraise them morally for their weak-willed actions (although it might be appropriate to appraise them for failing to cultivate appropriate levels of self-control).

Watson's argument against the commonsense notion of weakness of will is flawed. It assumes that exercising the capacity to act on our best judgment is, like many capacities, action-like: we must choose to exercise it. Hence, if we do not exercise it, it must either be because we chose not to or did not try hard enough. Yet Watson argues that neither works, and so we cannot think of weakness of will as an unexercised capacity. The fundamental problem with Watson's argument is its starting point: the assumption that exercising our capacity to act on our judgment about what is best is action-like, such that we must choose to exercise it. This assumption leads to a vicious infinite regress.

To see this, suppose that P judges that it is best to do A but experiences contrary motivation. Since P judges that it is best to do A, he judges that it is best to resist contrary motivation. In other words, he judges that it is best to exercise self-control to do A. Let "SC1" stand for P's capacity of self-control to do A. On Watson's assumption, exercising SC1 is a choice. How can P choose to exercise SC1? Because P judges that it is best to exercise SC1, his choosing to exercise SC1 requires P to exercise his capacity to act on his judgment about what is best in the face of contrary motivation. In other words, P must exercise self-control in order to exercise SC1. Let "SC2" stand for P's capacity of self-control to exercise SC1. On Watson's assumption, exercising SC2 is also a choice. Further, because P judges that it is best to exercise SC1, P judges that it is best to resist contrary motivation, and so P judges that it is best to exercise SC2. This means that, on Watson's assumption, P must exercise his capacity to act on his judgment about what is best – i.e., he must exercise self-control – in order to exercise SC2. Let "SC3" stand for P's capacity of self-control to exercise SC2. Exercising SC3 is also a choice, on Watson's assumption, and we clearly have a vicious infinite regress here: in order to choose to exercise the capacity of self-control at level n, you must exercise the capacity of self-control at level n+1.[9]

The solution is to recognize that exercising the capacity to act on our judgment about what is best is not action-like. It is not something that we choose to do. Other rational capacities are like this too. For example, we do not choose to exercise our capacity to make inferences.[10] Because Watson's starting point is flawed, he has not undermined the commonsense view of weakness of will as the agent's failure to exercise his capacity to act on his judgment about what is best. I take it then that we should accept the commonsense view and so accept my argument that we are attributionally responsible for Type 2 weak-willed actions.

[9] I thank ***… For a similar line of objection, see Kennett (2001), pp. 135-140 and162.

[10] Not only does it seem phenomenologically wrong that we choose to exercise our capacity to make inferences, but that assumption also leads to an analogous vicious infinite regress (Carroll 1895).

4. The Depth of the Weakness of Will Problem

As we saw in Section 2, Watson can broaden his view of attributability to overcome Fischer's objection. On this broader view, an agent acts freely in the sense relevant to attributability if and only if he acts on what he values (rather than on what he most values). Yet, as we saw in Section 3, this view incorrectly says that agents are not attributionally responsible for Type 2 weak-willed actions, or actions in which an agent weakly acts in pursuit of an end that he does not value at all.

The fact that the broader view still cannot account for weakness of will may seem to reinforce Fischer's conclusion that views of attributability that lack a normative component altogether – such as Harry Frankfurt's – are right not to incorporate one. In fact, Fischer says that "… the problem of weakness of will does not threaten Frankfurt's 'minimalist' hierarchical approach to giving an account of the sort of freedom implicated in moral responsibility" (Fischer, 2010, p. 314).

Fischer is wrong on this point. Non-normative views of attributability like Frankfurt's also succumb to the weakness of will problem. To see this, examine Frankfurt's (1971/2003) original view. On this view, an agent acts freely (in the sense relevant to attributability) if and only if he *identifies* with the first-order desire that leads to his action. An agent identifies with a first-order desire *x* when he has a "higher-order volition" with respect to *x* – i.e., a higher-order desire that *x* move him all the way to action. This view cannot account for the fact that we are attributionally responsible for weakly acting on a desire with which we do not identify. As Frankfurt admits, when an agent identifies with one desire but not with a conflicting one, "the conflict between the two desires [may] remain as virulent as before" (Frankfurt 1987/1988, p. 172). Thus the "outlaw" desire, the one with which he does not identify, may have significant motivational force and so may lead to action. For example, we can assume that Watson's squash player does not identify with his desire to strike his opponent with his racquet; yet it may so move him anyway. Further, an outlaw desire may move an agent to action even though he judges that it should not and has the capacity to resist it. So the squash player may weakly act on the desire to strike his opponent with his racquet despite not identifying with it, and as we have seen, he would be attributionally responsible for that

action. After all, it would be appropriate to evaluate him as "hot-headed" for it, which is an evaluation of him as an agent. Frankfurt's view fails to give this result.

Unlike Fischer, others have recognized that Frankfurt's view fails to account for weakness of will (e.g., Haji 2002). However, the general problem has not been recognized.[11] The general problem is this: because we are always attributionally responsible for weak-willed actions, we are attributionally responsible for weak-willed actions that result from failing to exercise the capacity relevant to attributability.

No current view of attributability accounts for this fact. These views differ amongst each other in what they claim the attributability-relevant capacity to be – for example, on the broader version of Watson's view, it is the capacity to act on your values; on Frankfurt's view, it is the capacity to act on a desire with which you identify; and on David Shoemaker's (2003) view, it is the capacity to act on your cares – yet they all claim that you must *exercise* the attributability-relevant capacity in order to be attributionally responsible for an action. In other words, they all accept the following general principle:

> Exercise of Capacity Principle (EC Principle): An agent is attributionally responsible for an action A if and only if A results from the exercise of his attributability-relevant capacity to do A.

Indeed, the EC Principle is so taken for granted that it is even incorporated into general descriptions of attributability that are neutral amongst particular views. For example, in linking attributability to agent-evaluations, Watson says: "[w]hen thought or behavior are *exercises* of what Dewey calls an agent's moral capacity [i.e., an agent's attributability-

[11] Sarah Buss (2012) claims that weakness of will creates a general problem for "endorsement theories of autonomy," but the problem that she identifies is not general enough. On her understanding of an endorsement theory, an action is autonomous if and only if the agent willingly endorses it. She cites Watson's original view and Frankfurt's view as examples of endorsement theories. Against these views, she rightly points out that weak-willed agents can act autonomously and so be responsible for their weak-willed actions, since "the fact that they repudiate what they are doing even as they do it does not free them from blame" (p. 656). Yet Buss does not recognize the depth of the problem. The broader version of Watson's view is not an endorsement theory on her understanding, since she makes clear that we willingly endorse at most one action in any situation, and so the broader view allows agents to be responsible for weakly acting against what they willingly endorse. As a result, the broader view evades the difficulty that Buss points out.

relevant capacity], they and their results are open to distinctive kinds of evaluation. These evaluations are inescapably evaluations of the agent…" (Watson 1996/2004, p. 233, italics mine). Although this passage technically refers only to the sufficiency condition of the EC Principle, Watson clearly intends the necessity condition as well. The quoted passage appears within a discussion of the nature of attributability, and there is no reason to think that he shifts from the nature of attributability to a sufficient but unnecessary condition of it. More strongly, in a later discussion of the relationship between attributability and control, Watson explicitly claims that not exercising the attributability-relevant capacity negates attributability. He says: "'I couldn't help it' negates responsibility [attributability], for example, only by indicating that the individual's behavior (or omission) was not after all an exercise of 'moral capacity'" (1996/2004, p. 234).

Although the EC Principle provides a sufficient condition for attributability, it does not provide a necessary one. As we have seen, we are attributionally responsible for all weak-willed actions, including weak-willed actions that result from failing to exercise the attributability-relevant capacity, since weak-willed actions express our moral identity. To illustrate the general problem, consider again Watson's squash player. Whatever the attributability-relevant capacity is, we can assume that the squash player has that capacity but fails to exercise it, resulting in his smashing his opponent in the face with his racquet. In this case, the EC Principle says that the squash player is not attributionally responsible for his weak-willed action because he does not exercise the attributability-relevant capacity. This is incorrect. He is attributionally responsible for his action, as it is appropriate to appraise him as hot-headed for it. Thus the EC Principle fails to provide a necessary condition for attributability.

Before presenting my solution, consider David Shoemaker's (2003) claim that his view of attributability accounts for weakness of will, which seems at odds with my claim that it does not. Shoemaker conceives of weakness of will as failing to act on something that you generally care about more – i.e., eating healthily – in favor of something that you care about more in your specific circumstances – i.e., enjoying desserts at a party. His view accounts for his conception of weakness of will, but I think that his conception is flawed. His conception is insufficient for weakness of will because it does not

differentiate a weak-willed dessert eater from one who judges that it is best to relax his healthy-eating habits just this once. More importantly, his conception is not necessary for weakness of will. Consider a person who cares about eating the desserts but even at the party still cares more about healthy eating. If he eats the desserts anyway in spite of having the capacity to resist, he is weak-willed. Similarly, a dessert eater who does not care at all about eating the desserts is, like Watson's squash player, weak-willed if he has the capacity to resist. Yet ultimately, nothing rides on the terminology. What matters for my purposes is that an agent is attributionally responsible for acting on an end that he does not care about at all when he has the capacity to act on his best judgment. Shoemaker's view cannot account for this point, whether or not we call such an agent "weak-willed."

5. The Lesson for Attributability: We Just Need to Have the Capacity

We have just seen that all views of attributability on offer are flawed because they accept the EC Principle. Although the EC Principle provides a sufficient condition for attributability, it wrongly claims that agents are not attributionally responsible for weak-willed actions that result from failing to exercise the attributability-relevant capacity. In this section, I argue that we can solve the weakness of will problem by replacing the EC Principle with a new principle, which fortunately any view of attributability can adopt. To arrive at it, first consider the following general principle as a replacement for the EC Principle:

> Initial Proposal: An agent is attributionally responsible for an action A if and only if he has the attributability-relevant capacity to perform A or avoid performing A.

Initial Proposal accounts for our data so far. An agent who exercises the attributability-relevant capacity to do A has that capacity, which explains why the EC Principle provides a sufficient condition for attributability. Further, Initial Proposal allows an agent to be attributionally responsible for weak-willed actions that result from

failing to exercise the attributability-relevant capacity. For example, it says that the squash player is attributionally responsible for smashing his opponent in the face, since he has the attributability-relevant capacity to avoid it.

The spirit of Initial Proposal is correct. Being attributionally responsible for an action is not fundamentally a matter of exercising the attributability-relevant capacity but of *having* the attributability-relevant capacity. Yet to be attributionally responsible for an action, it matters how the action comes about. Specifically, to be attributionally responsible for an action, the agent must perform it *because* he possesses the attributability-relevant capacity and either exercises it or fails to exercise it.

To see this, consider the following Frankfurt-style case.[12] Suppose that Susan wants Scott, a notorious bank robber, to rob Bank X. Susan, an excellent neuroscientist, has secretly implanted a chip in Scott's brain that gives her the power to detect whether he is about to decide to rob a bank and the power to force him to do so. Now suppose that Scott is deliberating about whether to rob Bank X. If he is about to decide to rob it, Susan will do nothing, and Scott will rob Bank X by exercising his attributability-relevant capacity, making him attributionally responsible for it. However, if Scott is about to decide not to rob Bank X, Susan will use the chip to force him to rob it. Now suppose that Scott is about to decide not to rob Bank X, and so Susan forces him to do so. Clearly, Scott is not attributionally responsible for robbing Bank X, as it does not express his moral identity when it comes about as a result of Susan's interference. Yet he has the attributability-relevant capacity to rob the bank, which he refrained from exercising, and so it might seem that Initial Proposal wrongly says that he is attributionally responsible for robbing Bank X.

The solution is to recognize that Scott's having the attributability-relevant capacity to rob Bank X is irrelevant to the explanation of his robbing Bank X. When Susan forces him to rob Bank X, she bypasses his attributability-relevant capacity, and so her interference, not his capacity, explains his action. The upshot is that, to be attributionally responsible for an action, an agent must simply have the attributability-relevant capacity – he need not exercise it – but his having this capacity must figure in the explanation of his action.

[12] Frankfurt (1969/2003) introduced this type of case.

To account for this point, we should modify Initial Proposal as follows:

> 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.

In both conditions 1) and 2), the agent's having the attributability-relevant capacity – and either exercising it (condition 1) or failing to exercise it (condition 2) – figures in the explanation of how action A comes about.

The HC Principle is correct. Condition 1) is the sufficient condition from the EC Principle, and so the HC Principle correctly says that we are attributionally responsible for actions that result from exercising our attributability-relevant capacity to do them. The HC Principle also accounts for cases like the squash player, which the EC Principle gets wrong. According to the HC Principle, the squash player is attributionally responsible for smashing his opponent in the face with his racquet, even though that action does not result from his exercising his attributability-relevant capacity, because he meets condition 2): his smashing his opponent in the face results from his failing to exercise his attributability-relevant capacity to avoid it. This correctly explains why the squash player is attributionally responsible for smashing his opponent in the face with his racquet.

Yet cases like the squash player are rare. On any plausible understanding of the attributability-relevant capacity, weak-willed actions typically result from an agent's exercising the attributability-relevant capacity to perform them. For example, weak-willed actions typically result from an agent's pursuing an end that he values (Type 1 weakness of will), since it is rare for an agent to act weakly in pursuit of an end that he does not value at all (Type 2 weakness of will). Hence, if the attributability-relevant capacity is the agent's capacity to act on his values, weak-willed actions typically result from the agent's exercising his attributability-relevant capacity to perform them. Similarly, weak-willed actions usually result from an agent's cares and from a desire with which he identifies. Therefore, if the attributability-relevant capacity is the agent's

capacity to act on his cares or his capacity to act on a desire with which he identifies, weak-willed actions typically result from the agent's exercising his attributability-relevant capacity to perform them. The upshot is that condition 1) of the HC Principle says that agents are attributionally responsible for their weak-willed actions in typical cases of weakness of will. Does this mean that condition 2) is irrelevant to typical cases of weakness of will?

It may initially seem so. In that case, condition 2) would not account for attributability for weakness of will as such. Condition 2) may even seem insignificant, as it would only come into play in rare cases of weakness of will. Of course, condition 2) would still be necessary to account for attributability in those rare cases, and so the HC Principle would still be correct. But the HC Principle's superiority over the EC Principle might seem marginal, like I am pointing out a technicality.

This line of reasoning is flawed. First, condition 2) would be significant even if it only came into play in rare cases of weakness of will. After all, it is important for understanding the nature of attributability that we can be attributionally responsible for actions that do not result from exercising the attributability-relevant capacity. I will have more to say about this below. Second, contrary to initial appearances, condition 2) is crucial for accounting for attributability for weakness of will as such. Here is why: weakness of will is an omission – the agent's failure to act on his best judgment – that results from his failure to exercise his attributability-relevant capacity to avoid it. Thus we need condition 2) to say that an agent is attributionally responsible for his weakness of will.

To see this second point, recall that a weak-willed agent judges that it is best to do A and has the capacity to do A, but he does something else, B, instead. What *makes* the agent weak-willed is not his doing B or even his failing to do A. After all, an agent can do B or fail to do A without being weak-willed, such as when he judges that doing B or failing to do A is best. What makes the agent weak-willed is his failure to act on his best judgment (when he has the capacity to act on it). To illustrate, recall the CEO who judges that he should tell his shareholders the truth but weakly lies to them. The CEO is not weak-willed because he lies to his shareholders or because he fails to tell them the truth. After all, a more ruthless CEO would lie because he judges it best to do so.

Rather, the weak-willed CEO is weak-willed because he fails to act on his best judgment (when he has the capacity to act on it). The upshot is that the defining omission of weakness of will is this: an agent's not acting on his judgment about what is best.

Only condition 2) says that weak-willed agents are attributionally responsible for not acting on their best judgment. To make this clear, consider the HC Principle written explicitly for omissions:

> HC Principle for omissions: An agent is attributionally responsible for not doing A if and only if 1) his not doing A results from the exercise of his attributability-relevant capacity to not do A or 2) his not doing A results from his failure to exercise his attributability-relevant capacity to do A.

Next notice that a weak-willed agent's capacity to act on his best judgment is attributability-relevant. After all, agents are attributionally responsible for acting on their best judgments. Thus a weak-willed agent has the attributability-relevant capacity to act on his best judgment, and his not acting on it results from his failure to exercise his attributability-relevant capacity to act on it. Hence the weak-willed agent meets condition 2) – but not condition 1) – with respect to not acting on his best judgment. Thus condition 2) is needed to say that weak-willed agents are attributionally responsible for not acting on their best judgments, and so condition 2) accounts for attributability for weakness of will as such.

Of course, in a particular case, a weak-willed agent judges that it is best to do A but does B instead, and so his failure to act on his best judgment is constituted by his not doing A, which in turn is constituted by his doing B. Hence, a weak-willed agent must be attributionally responsible for not doing A and for doing B. Only condition 2) says that he is attributionally responsible for not doing A: his not doing A results from his failure to exercise his attributability-relevant capacity to do A. What about B? Here is where things get more complicated.

In the rare cases in which the weak-willed agent does not exercise his attributability-relevant capacity to do B, only condition 2) says that he is attributionally responsible for B. As we have seen, the squash player is attributionally responsible for

smashing his opponent in the face with his racquet just because he fails to exercise his attributability-relevant capacity to avoid doing so. Of course, in this case, his doing B is equivalent to his not doing A: he judges that he should not smash his opponent in the face, and so his failure to act on his best judgment is equivalent to his smashing his opponent in the face. However, even when B is not equivalent to failing to do A, the same point holds. Suppose that the squash player judges that he should walk away, but he instead smashes his opponent in the face with his racquet despite having the capacity to walk away. Assuming that he does not exercise the attributability-relevant capacity to smash his opponent in the face, he is still attributionally responsible for doing so because he meets condition 2) with respect to it.

Yet weak-willed agents typically exercise their attributability-relevant capacity to do B. In these cases, condition 1) and condition 2) *both* say that the agent is attributionally responsible for B: B results from his exercising his attributability-relevant capacity to do B, and B results from his failing to exercise his attributability-relevant capacity to avoid doing B. But recall that B is a weak-willed action because it constitutes the agent's failure to act on his best judgment. Thus the agent is attributionally responsible for B *as a weak-willed action* because he meets condition 2) with respect to it. Hence even when an agent exercises his attributability-relevant capacity to B, condition 2) accounts for the fact that he is attributionally responsible for B as a weak-willed action.

I have just shown that only condition 2) accounts for attributability for weakness of will. Importantly, this means that any view of attributability can account for weakness of will by embracing the HC Principle. Return to Watson's original view: an agent acts freely in the sense required for attributability if and only if he acts on what he most values. Above, we saw that this view cannot account for Type 1 weakness of will, like in the CEO case. Yet by adding condition 2), Watson's original view would correctly claim that the CEO is attributionally responsible for lying to his shareholders, since it results from his failure to exercise his capacity to act on what he most values. However, a significant problem remains: the CEO is also attributionally responsible for lying to his shareholders because it results from his acting on his values. Thus, although Watson's

original view could account for weakness of will by adding condition 2), it proposes too robust of an attributability-relevant capacity.

I have shown, then, that any view of attributability must embrace condition 2) to account for weakness of will. Further, no other condition is needed, as conditions 1) and 2) cover both the exercise and the failure to exercise the attributability-relevant capacity. Hence we should reject the EC Principle in favor of the HC Principle. This makes sense. An agent's failure to avoid an action when he has the attributability-relevant capacity to avoid it expresses his moral identity, making him attributionally responsible for that failure.

Before concluding, we should consider how shifting from the EC Principle to the HC Principle affects our understanding of the nature of attributability. Sometimes attributable actions are generically thought of as actions that the agent endorses or stands behind or that result from desires that the agent endorses or stands behind. This way of understanding attributability only works with the EC Principle. After all, the weak-willed squash player does not endorse or stand behind his smashing his opponent in the face with his racquet, nor does he endorse or stand behind the desire for that. You might then wonder how the HC Principle changes this understanding.[13]

On my view, we use language such as "endorsing" or "standing behind" an action or desire to convey that attributable actions express an agent's moral identity. I have therefore worked with the idea that attributable actions are those that express an agent's moral identity, and shifting from the EC Principle to the HC Principle does not affect that understanding. Indeed, my argument for the HC Principle relies on it. Hence I see no problem in jettisoning the language of endorsement if it is misleading in certain cases like the squash player. Yet I think that we could continue to use the language of endorsement and still take account of the HC Principle as follows: attributable actions are those actions that result from the agent's capacity to act in a way that he endorses (or to act on a desire that he endorses) – either from exercising or failing to exercise that capacity.

But why is an action attributable to an agent just in virtue of resulting from his attributability-relevant capacity, even if that capacity is not exercised? On my view, the answer is this: when an action results from the agent's attributability-relevant capacity,

[13] Thank you to an anonymous referee for raising this issue.

he has control over the fact that he performs it. An agent clearly has control over the fact that he performs action A when A results from his exercising his attributability-relevant capacity to perform A. Yet an agent also has control over the fact that he performs A when A results from his failing to exercise his attributability-relevant capacity to avoid A, and this is precisely because he could have avoided A. Recall my above analogy of the drowning man and two bystanders. Although neither bystander attempts a rescue, the one who can swim has control over the fact that the man drowns precisely because he can save him. Similarly, the squash player has control over the fact that he smashes his opponent in the face with his racquet precisely because he can avoid it by exercising his attributability-relevant capacity to avoid it. He is not like an agent who is compelled to do it, who would lack such control.

My argument here generalizes one that I made above. Using the drowning man and bystanders analogy, I argued that an agent who weakly acts on a desire for an object that he does not value is not a mere bystander to his action, since he could have acted on his values. The general point is this: an agent who does A because he fails to exercise the attributability-relevant capacity to avoid doing A is not a mere bystander to his doing A, since he could have exercised the attributability-relevant capacity to avoid doing A. Saying that he is not a mere bystander to his doing A is just another way of saying that he has control over the fact that he does A. Hence, I think that the HC Principle supports the idea that an agent is attributionally responsible for an action if and only if he has control over the fact that he performs it.

This interpretative suggestion needs to be fleshed out. Yet whether or not it withstands scrutiny, my argument for the HC Principle stands, and so we must wrestle with why actions that result from failing to exercise the attributability-relevant capacity express an agent's moral identity.

6. Conclusion

Although particular views of attributability have been criticized for failing to account for weakness of will – in this paper, I highlighted Fischer's criticism of value-

added views of attributability – it has not previously been recognized how deep the problem is. As I argued, all current views of attributability fail to account for weakness of will because they all accept the following general principle:

> EC Principle: An agent is attributionally responsible for an action A if and only if A results from the exercise of his attributability-relevant capacity to do A.

The EC Principle is wrong because, as weakness of will cases show, we are also attributionally responsible for actions that result from *failing* to exercise the attributability-relevant capacity to avoid them. Hence, to account for weakness of will, we must adopt the following general principle instead:

> 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 the failure to exercise his attributability-relevant capacity to avoid doing A.

I first argued for the HC Principle by noting that some weak-willed actions do not result from the agent's exercising his attributability-relevant capacity to perform them. An agent is attributionally responsible for these weak-willed actions, I argued, because they express his moral identity, but the EC Principle fails to give this result. The HC Principle gets these cases right, thanks to condition 2). Yet, importantly, the HC Principle is needed to account for attributability for weakness of will as such. As I demonstrated, only condition 2) correctly says that a weak-willed agent is attributionally responsible for failing to act on his judgment about what is best – the defining omission of weakness of will. Condition 2) also rightly claims that a weak-willed agent is attributionally responsible for that which constitutes his failure to act on his best judgment: his not doing A and his doing B, when he judges that he should do A but does B instead. Although B typically results from the agent's exercising his attributability-relevant capacity, the agent is attributionally responsible for B *as a weak-willed action*

because condition 2) holds. Hence we need condition 2), and so the HC Principle, to account for attributability for weakness of will.

Because the HC Principle is correct, attributability is not a matter of exercising the attributability-relevant capacity, as current views of attributability assume. Rather, it is a matter of having that capacity – so long as having that capacity figures in the explanation of the action. Fortunately for current views, any view can adopt this reconceptualization of attributability.

Acknowledgments

I thank Lara Buchak, Gideon Rosen, and Michael Smith for valuable discussions and comments, and anonymous referees for helpful comments, on earlier versions of this paper.

References

Arpaly, Nomy. (2003). *Unprincipled Virtue,* Oxford University Press.

Buss, Sarah. (2012). Autonomous Action: Self-Determination in the Passive Mode. *Ethics,* 122(4), 647-691.

Carroll, Lewis. (1895). What the Tortoise Said to Achilles. *Mind*, 4(14), 278-80.

Fischer, John Martin. (2010). Responsibility and Autonomy. In T. O'Connor and C. Sandis (eds.) *A Companion to the Philosophy of Action* (pp. 309-316). Blackwell Press.

Frankfurt, Harry. (2003.) Alternate Possibilities and Moral Responsibility. In Gary Watson (ed.) *Free Will* (pp. 167-176). Oxford: Oxford University Press, 2003. (Original work published 1969)

Frankfurt, Harry. (2003). Freedom of the Will and the Concept of a Person. In Gary Watson (ed.) *Free Will* (pp. 322-336). Oxford: Oxford University Press. (Original work published 1971)

Frankfurt, Harry. (1988). Identification and Wholeheartedness. In his *The Importance of What We Care About* (pp. 159-176). Cambridge University Press. (Original work published 1987)

Haji, Ishtiyaque. (2002). Compatibilist Views of Freedom and Responsibility. In R. Kane (ed) *The Oxford Handbook of Free Will* (pp. 202-228). New York: Oxford University Press.

Kennett, Jeanette. (2001). *Agency and Responsibility*, Oxford University Press.

Mitchell-Yellin, Benjamin. (2014). In Defense of the Platonic Model: A Reply to Buss. *Ethics*, 124 (2), 342-357.

Shoemaker, David. (2003). Caring, Identification, and Agency. *Ethics* 114(1), 88-118.

Smith, Michael. (2003). 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.

Vihvelin, Kadri. (1994). Are Drug Addicts Unfree? In S. Luper-Foy and C. Brown (eds.) *Drugs, Morality and the Law* (pp. 51-78). New York: Garland.

Watson, Gary. (2004). Free Action and Free Will. In his *Agency and Answerability* (pp. 161-196). Oxford University Press. (Original work published 1987)

Watson, Gary. (2004). Free Agency. In his *Agency and Answerability* (pp. 13-32). Oxford University Press. (Original work published 1975)

Watson, Gary. (2004). Skepticism About Weakness of Will. In his *Agency and Answerability* (pp. 33-58). Oxford University Press. (Original work published 1977)

Watson, Gary. (2004). Two Faces of Responsibility. In his *Agency and Answerability* (pp. 260-288). Oxford University Press. (Original work published 1996)