NON-PSYCHOLOGICAL WEAKNESS OF WILL: SELF-CONTROL, STEREOTYPES, AND CONSEQUENCES

Forthcoming in *Synthese*

Mathieu Doucet and John Turri

University of Waterloo

1. Introduction

Philosophical discussion of weakness of will has long treated it as a thoroughly psychological phenomenon. In recent years, that debate focused on the question of precisely what the phenomenon is, and whether actions contrary to an agent's better *judgments* (Mele 1987, 2010) or her *intentions* (Holton 1999, 2009) are more central to the concept. An important part of this debate has been a disagreement about the ordinary notion of weakness of will: Richard Holton (1999) set the debate off by claiming that, in emphasizing the violation of judgments rather than intentions, philosophers have failed to describe what ordinary people mean by weakness of will. More recently, something of a consensus has emerged, as defenders of both the judgment- and intention-centered accounts have now reported on experiments that suggest that the folk concept includes actions that violate *either* kind of commitment-entailing mental state (Mele 2012, May & Holton 2012).

In this paper, we show that the psychological focus on violated commitments –- whether judgments, intentions, or both –- has been too narrow. Recent research has show that weakness of will attributions are sensitive to the moral valence of the action, a non-psychological factor (May & Holton 2012, Sousa & Mauro forthcoming). We extend these findings beyond moral valence, while also demonstrating that the influence of non-psychological factors cannot be explained by appeal to the attribution of implicit psychological states. We begin by showing that

a sizable minority of people attribute weakness of will even in the absence of a violated commitment (Experiment 1). We then show that weakness of will attributions are sensitive to two important, and new, non-psychological factors. First, for actions stereotypically associated with weakness of will, the *absence* of certain commitments often triggers weakness of will attributions (Experiments 2-4). Second, the quality of an action's outcome affects the extent to which an agent is viewed as weak-willed: actions with bad consequences are more likely to be viewed as weak-willed (Experiment 5). Prior work on the ordinary notion of weakness of will has therefore told only half the story. Whereas prior accounts seem to have admirably captured the psychological side of the ordinary notion, they have neglected its non-psychological side. The ordinary concept of weakness of will is sensitive to two non-psychological factors and is thus much broader than philosophers have thus far imagined. We conclude by suggesting a two-tier model of weakness of will as a failure of self-control that unifies our findings with prior philosophical models.

2. Experiment 1: Weakness of will without commitment violations?

Until recently, one of the central debates about weakness of will was whether the violation of a judgment or an intention is more central to the concept. Mele calls both mental states "practical commitments": a judgment is an evaluative practical commitment, and a resolution or intention is an executive practical commitment (Mele 2012, p. 16). Cases that feature both are clear instances of weakness of will. Cases that feature the violation of one but not the other, however, are less clear. On Mele's account, cases that feature the violation of a judgment but not an intention are genuine instances of weakness of will, while on Holton's, they are not (Mele 1987,

2010; Holton 2009).[1] Mele, then, offers a disjunctive account: weakness of will involves the violation of either kind of practical commitment. Holton, by contrast, offered a more restrictive account: weakness of will involves the violation of resolutions (a form of intention), and it is only when judgment and resolution are in agreement that judgment violations count as weak-willed.

Recent experimental work by Holton and co-author Josh May has moved toward Mele's view that both forms of commitment violation are part of the folk concept. Some differences remain, however. May and Holton argue that the ordinary notion is a cluster concept rather than a disjunctive one. On their view, cases that feature both commitment violations are more central, while cases that feature one but not the other are equally marginal (May and Holton 2012). Even more recently, James Beebe has shown that the violation of *either* a judgment *or* a resolution is sufficient to generate folk ascriptions of weakness of will, which offers support for Mele's disjunctive account (2013, under review).

The emerging consensus, then, is that weakness of will involves the violation of practical commitments. This view, however, has emerged out of a debate about *which* commitment violations are more central to the ordinary notion of weakness of will. This means that less attention has been paid to whether such violations are in fact essential to the folk concept. In order to establish that the concept essentially involves practical commitment violation, we need to establish that in cases that lack such violations, people will not make weakness of will attributions. Mele, however, does not consider this question: all of his experiments involve

[1] There is a further difference between the two views that is not explored in the recent debate between Mele and May & Holton. On Holton's model, weakness of will frequently involves irrationally *revising* one's commitments so that at the time of action there is no inconsistency between judgment, intention, and action (2009). Mele, by contrast, emphasizes that core cases involve an agent who is consciously aware, at the time of action, that she is violating her own commitments (2012).

commitment violations. So nothing in his experiments establish that practical commitment violations are essential to the folk concept of weakness of will. While May and Holton do ask about such cases, their discussion of them is limited.

Experiment 1 corrects this oversight by testing whether practical commitment violation is in fact a central element of the folk concept. We contrast a case with a clear violation with one that clearly lacks such a violation. If the folk concept really is just practical commitment violation, then we should see frequent attributions of weakness of will in cases that feature such violations and far fewer attributions in cases that lack them.

2.1 Method

Participants (N = 65)[2] were randomly assigned to two conditions: Violation and No Violation. Each participant read a single story. Here are the stories used in the two conditions (Violation/No Violation manipulations in brackets):

> Peter is a very [conventional/unconventional] person. Doing well in his classes [is very important to him/has never mattered to him], and he has made a commitment to himself to [do whatever it takes to achieve academic success/not get caught up in the myth of academic success]. In particular, he has made a commitment to himself to [study hard for every test/never study for a test]. ¶ There is a major test in Peter's science class on Friday. On Thursday

[2] Twenty-nine female, aged 18–62, mean age = 30.4, 97% reporting native competence in English. Data from four participants who had participated in a previous study were excluded from the analysis. As with the experiments reported below, participants were recruited using Amazon Mechanical Turk and compensated $.30 for approximately 2–3 minutes of their time. Participation was restricted to United States residents. They filled out a brief demographic survey after testing.

night, Peter is invited to a party. He decides to go to the party instead of studying for the test. He stays out very late. Because he didn't study, he fails the test on Friday.

After reading the story, participants were asked to rate their agreement or disagreement with three statements:

1. Peter has an appropriate attitude toward his studies.
2. By going to the party, Peter broke the commitment he made to himself.
3. By going to the party, Peter displays weakness of will.

Responses were collected on a standard Likert scale, 1 (=Strongly disagree) through 7 (=Strongly agree). Question 2 was a manipulation check to ensure that participants recognized whether Peter had broken a commitment. We asked question 1 to check whether participant evaluation of Peter's attitude mediated the effect of condition on response to the weakness of will question.

2.2 Results

The manipulation was effective. Participants in Violation overwhelmingly agreed that Peter broke his commitment (M = 6.16, SD = 1.29), whereas participants in No Violation overwhelmingly disagreed (M = 1.53, SD = 1.11), independent samples t-test, $t(63) = -15.55$, $p < .001$, MD = -4.63, d = 3.92[3]

[3] We report effect sizes using Cohen's d (Cohen 1988), which is a standard statistic for interpreting effect size when comparing mean differences ("MD" abbreviates "mean

There was an effect of condition on agreement with the weakness of will attribution, with scores significantly higher in the Violation condition (M = 5.42, SD = 1.21) than in the No Violation condition (M = 3.91, SD = 2.18), independent samples t-test, $t(52.36) = -3.49$, $p = .001$, MD = 1.51, d = 0.96. The mean score in No Violation did not differ from the neutral midpoint (= 4), one-sample t-test, $t(33) = -.236$, $p = .815$, n.s., whereas it was significantly above midpoint in Violation, $t(30) = 6.56$, $p < .001$.

A multiple regression analysis revealed that participant evaluation of Peter's attitude did not mediate the effect of condition on whether Peter was viewed as weak-willed. Entering response to the weakness of will statement as the outcome variable, and entering the independent variable (No Violation/Violation) and response to the attitude statement as predictors, the independent variable was significantly predictive, $t(62) = 3.59$, Beta = 0.495, $p < .001$, whereas response to the attitude statement was not, $t(62)$, Beta = -0.182, $p = .191$, n.s.

2.3 Discussion

These results support the view that the violation of commitments significantly affects weakness of will attributions and is central to that concept. The lack of such a violation significantly depresses weakness of will attributions. Nevertheless, the results from the No Violation condition are more than a little surprising. Overall, mean response to the weakness of will statement was neutral. More revealing is the distribution of responses (see Figure 1). Whereas in the Violation condition almost everyone agreed at least somewhat with the weakness of will statement, in the No Violation condition there was a striking bimodal distribution of responses. In particular, 14 of 33 participants (over 42%) agreed to some extent that Peter displayed

---

difference"). We follow Ellis (2010) for interpreting the magnitude of effect sizes. For d, values greater than or equal to .80 are large, greater than or equal to .50 but less than .80 are medium, and greater than or equal to .30 but less than .50 are small.

weakness of will by not studying for the test. Their agreement came despite the fact that Peter was overwhelmingly viewed as having *not* broken a commitment. Even among the 14 who agreed that Peter displayed weakness of will, the mean response to the statement that Peter had broken a commitment was 1.64, and the mode response (10 of the 14) was to "strongly disagree." A sizable minority of participants, then, thought Peter displayed weakness of will despite recognizing that he had not broken any commitment. These results are consistent with findings by Sousa and Mauro and Sousa (forthcoming). They found that a significant number of participants attributed weakness of will in cases that featured seriously immoral actions (e.g. assassination for money), even when those actions resulted from immoral resolutions.[4] By contrast, the present findings pertain to an action that is unlikely to be viewed as seriously immoral.

The surprising findings from the No Violation call out for explanation: after all, the philosophical consensus has been that the folk concept of weakness of will just is commitment violation. If there is a tendency to apply the concept in cases that lack such violations, then this opens up the possibility that the folk concept is broader than philosophers have thus far assumed.

[4] Fifty percent attributed weakness of will to a professional assassin who entertained doubts about killing, decided that the best thing to do was to continue killing, resolved to do so, and then killed. By contrast, only 15% attributed strength of will.

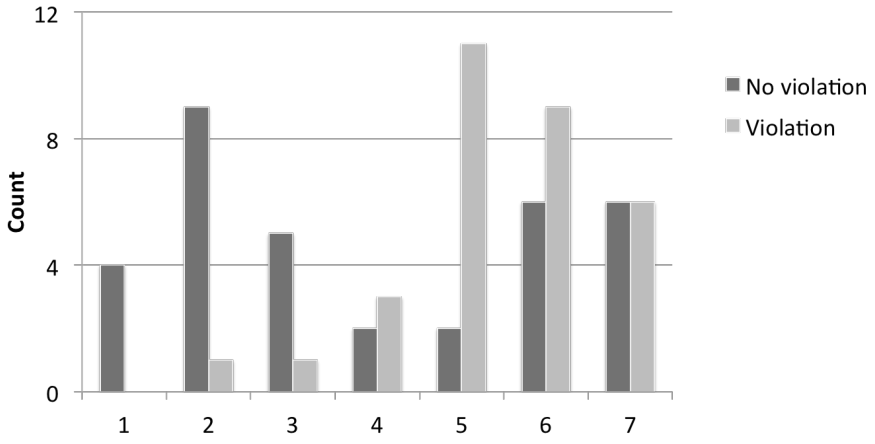


Fig. 1. Experiment 2: Distribution of responses to the weakness of will statement across conditions, 1 = Strongly disagree, 7 = Strongly agree.

One possible explanation for this surprising finding, inconsistent with prior theorizing on weakness of will, is that the folk concept counts *the failure to form some commitments* as itself weak-willed. More specifically, this tends to be true of conduct stereotypically associated with weakness of will. On this view, aside from the violation of practical commitments, in stereotypical cases *the absence of* certain commitments can directly trigger attributions of weakness of will. The tendency for philosophical discussions of weakness of will to concentrate on violations in stereotypical cases — which goes all the way back to Aristotle, who restricted *akrasia* proper to certain specific bodily pleasures (1999, III.10) — has perhaps obscured this point.

Another explanation for this surprising finding, consistent with prior theorizing, is that participants in the No Violation condition *implicitly attributed* a relevant commitment to the agent, and this commitment is broken in the story. This is how Sousa and Mauro (forthcoming) explain their finding. They suggest that, in cases featuring immoral resolutions and immoral

actions, participants reinterpret the content of the agent's judgment and assume that the person who acts immorally knows "deep inside" that his action is morally wrong.

Why would people implicitly attribute such a commitment? A long line of research in social psychology on "the false consensus effect" has shown that people overestimate the extent to which others share their views (Ross, Greene & House 1977; Marks & Miller 1987). It's very likely that many of our participants are themselves committed to academic success and, when the time comes, studying hard for an exam rather than staying out late at a party. Relatedly, a separate line of research suggests that people tend to view others as having a "true" or "deep" self (Johnson, Robinson, & Mitchell 2004) and that we tend to attribute attitudes and feelings to others' deep selves that correspond to our own views about what is good (Newman, Knobe, & Bloom, 2014). Such tendencies could lead many participants to implicitly attribute similar attitudes to the protagonist, including a commitment to academic excellence and studying for exams. More generally, stereotypical examples of weak-willed behavior tend to involve the violation commitments that many people share, so this explanation could generalize to other examples that elicited responses similar to those observed in this experiment.

So we have two competing explanations of the results. The first posits a previously unrecognized non-psychological factor that *directly* triggers weakness of will attributions in stereotypical cases.[5] The second posits that the effect of stereotypes is *indirect* and mediated by implicit commitment attributions, in keeping with participants' own evaluations and commitments, and which the agent violates. The first explanation represents a break with the current view that the folk concept centers on commitment violations, while the second is

[5] As will become clear below, to call such a factor 'non-psychological' simply means that weakness of will is not identified with the presence of any particular psychological process or state such as the violation of a commitment or the presence of particular judgment or intention. It does not mean that weakness of will does not involve intentional actions brought about by psychological processes.

consistent with that view. The next experiment begins testing between these two competing explanations.

3. Experiment 2: Stereotypical cases and implicit inferences

Frequent attribution of weakness of will in stereotypical cases lacking explicit commitment violations does not by itself show that the folk concept includes non-psychological elements. For stereotypical cases might be more likely to trigger inferences about the existence of background or implicit practical commitments. That is, participants might commonly assume that most people have a commitment to acting morally, or to getting good grades, or to not eating or drinking to excess. When confronted with cases in which people act immorally, or get bad grades, or drink to excess, participants might therefore be more likely to infer the existence of a violated commitment than they would be in less stereotypical cases. For example, participants might be more likely to infer that an agent *knew* it was bad for him to drink to excess[6] than when it comes to, say, excessively adding cards to a teetering house of cards. Similarly, participants might be more likely to infer that an agent knew it was bad for her to eat to excess than when it comes to, say, continuing to ride a ferris wheel excessively. The attribution of the relevant commitments could then explain why many participants view the agent as weak-willed in stereotypical cases, even when the agent does not explicitly have the relevant commitment.

This proposal generates testable predictions. Suppose we find that, other things being equal, participants are more likely to attribute weakness of will to agents lacking the relevant commitment in stereotypical cases than in non-stereotypical ones. That is, suppose there is a *stereotype effect* on weakness of will attributions. The proposal predicts that the stereotype effect

[6] Again we note that cases like this do not involve seriously immoral actions. Indeed, they arguably don't involve immorality at all.

will be indirect and mediated by participants' willingness to attribute practical commitments to the agent will mediate the stereotype effect. The present experiment tests whether this is indeed the case.

3.1 Method

Participants (N = 123)[7] were randomly assigned to one of four conditions in a 2 (Cover Story: Beth/Jesse) x 2 (Type: Stereotypical/Nonstereotypical) between-subjects design. Each participant read a single story. We used two different cover stories to ensure that results weren't driven by superficial features associated with any one form of conduct. Here are the stories (Stereotypical/Nonstereotypical variations in brackets):

> *Beth*: Beth went to a [restaurant/carnival] yesterday. At a certain point, it was evident to her that if she continued [eating the food being served, she would gain weight/riding the ferris wheel, she would not have enough money to go bungee jumping]. Beth continued [eating the food being served/riding the ferris wheel].

> *Jesse*: Jesse [went to a bar/was making a house of cards] last night. At a certain point, it was evident to him that if he continued [drinking, he would end up very drunk/adding cards, the house would collapse]. Jesse continued [drinking/adding cards].

Participants were then asked to rate their agreement or disagreement with three statements:

[7] Fifty-five female, aged 18–64, mean age = 32, 94% reporting native competence in English.

1. It was bad for Beth to continue eating/riding the ferris wheel.

   It was bad for Jesse to continue drinking/adding cards.

2. Beth knew it was bad for her to continue eating/riding the ferris wheel.

   Jesse knew it was bad for him to continue drinking/adding cards.

3. By continuing to eat/ride the ferris wheel, Beth exhibited weakness of will.

   By continuing to drink/add cards, Jesse exhibited weakness of will.

Responses were collected on a standard 7-point Likert scale, 1 (=Strongly disagree) to 7 (=Strongly agree).

3.2. Results

A preliminary analysis of variance revealed that Cover Story entered into both main and interaction effects. So we will analyze the results for each story separately, beginning with Beth.

For Beth's story, there was an effect of Type on response to whether it was bad for Beth to continue engaging in the activity, with mean response significantly higher in the Stereotypical condition (M = 4.47, SD = 1.61) than in the Nonstereotypical condition (M = 3.35, SD = 1.56), independent samples t-test, $t(59) = -2.736$, $p = .008$, MD = 1.12, d = 0.71. We also observed an effect of Type on response to whether Beth knew it was bad for her to continue engaging in the activity, with mean response significantly higher in the Stereotypical condition (M = 5.20, SD = 1.42) than in the Nonstereotypical condition (M = 3.61, SD = 1.71), independent samples t-test, $t(59) = -3.937$, $p < .001$, MD = -1.59, d = 1.03. Mean response in the Stereotype condition was significantly above the neutral midpoint (= 4), one-sample t-test, $t(29) = 4.616$, $p < .001$, but it

didn't differ from midpoint in the Nonstereotypical condition, $t(30) = -1.263$, $p = .216$, n.s. [8] Finally, agreement with the weakness of will statement was significantly higher in the Stereotypical condition (M = 4.63, SD = 1.87) than in the Nonstereotypical condition (M = 3.55, SD = 1.89), independent samples t-test, $t(59) = -2.253$, $p = .028$, MD = 1.08, d = 0.58. Attribution of weakness of will was trending above the midpoint in the Stereotypical condition, one sample t-test, $t(29) = 1.86$, $p = .073$, but was insignificantly lower than the midpoint in the Nonstereotypical condition, $t(30) = -1.327$, $p = .194$, n.s.

To test the mediating role of evaluative beliefs and the attribution of evaluative knowledge to the agent, we conducted a bootstrap multiple-mediators analysis (Hayes 2013) with condition as the independent variable, agreement with the weakness of will statement as the dependent variable, and agreement with the evaluative claim and knowledge attribution as potential mediators. This analysis indicated that knowledge attribution did not mediate the effect of condition, 95% confidence interval for an indirect effect = -0.23 to 1.14; but participant evaluation of the activity did mediate the effect of condition, 95% confidence interval for an indirect effect = 0.19 to 1.61. Moreover, the mediation was complete, 95% confidence interval for a direct effect condition = -0.76 to 0.87 (see Figure 2).

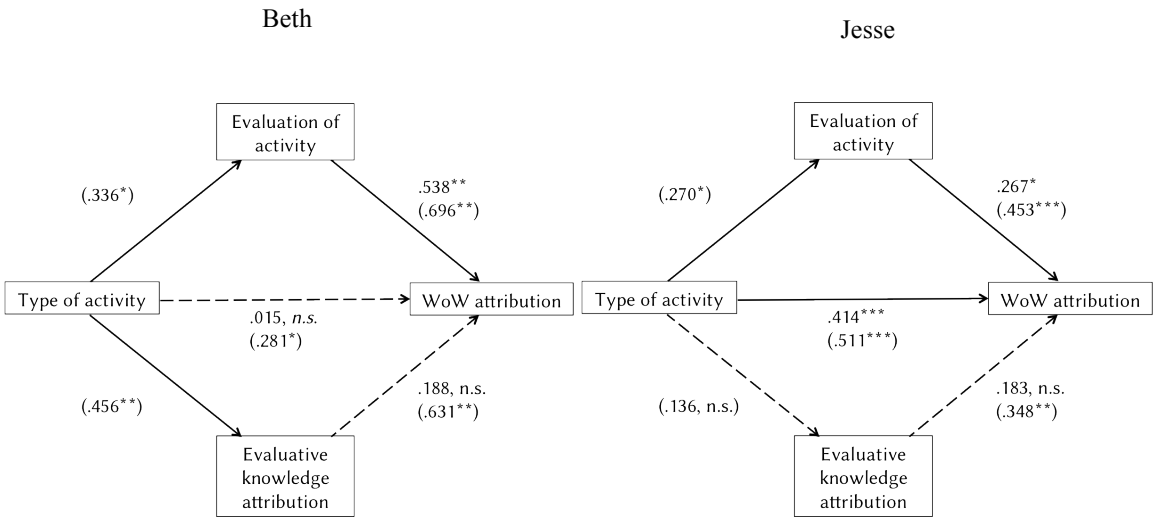
For Jesse's story, there was an effect of Type on response to whether it was bad for Jesse to continue engaging in the activity, with mean response significantly higher in the Stereotypical condition (M = 5.13, SD = 1.46) than in the Nonstereotypical condition (M = 4.39, SD = 1.23), independent samples t-test, $t(60) = -2.169$, $p = .034$, MD = 0.74, d = 0.56. There was no effect of Type on response to whether Jesse knew it was bad for him to continue engaging in the activity: Stereotypical, M = 5.65, SD = 1.17; Nonstereotypical, M = 5.29, SD = 1.44; $t(60) = -1.064$, $p =$

[8] Interestingly, participants in the Stereotypical condition more strongly agreed that *Beth knew it was bad for her to continue eating* than that *it was bad for her to continue eating*, one-sample t-test, $t(29) = -2.80$, $p = .009$, MD = 0.86, test proportion = 4.47.

.292, n.s. Mean knowledge attribution in both conditions was significantly above the neutral midpoint (= 4), one sample t-tests, *p*s < .001.[9] Finally, agreement with the weakness of will statement was significantly higher in the Stereotypical condition (M = 5.10, SD = 1.42) than in the Nonstereotypical condition (M = 3.39, SD = 1.50), independent samples t-test, $t(60) = -4.607$, $p < .001$, MD = 1.71, d = 1.19. Attribution of weakness of will was significantly above midpoint (=4) in the Stereotypical condition, one sample t-test, $t(30) = 4.293$, $p < .001$, and below midpoint in the Nonstereotypical condition, $t(30) = -2.28$, $p = .030$.

We again conducted the same bootstrap multiple-mediators analysis with condition as the independent variable, agreement with the weakness of will statement as the dependent variable, and agreement with the evaluative claim and knowledge attribution as potential mediators. This time neither candidate mediated the effect of condition on weakness of will attribution: 95% confidence interval for an indirect effect through knowledge attribution = -0.04 to 0.45; 95% confidence inter for an indirect effect through participant evaluation of the activity = 0.00 to 0.29. There was a substantial direct effect of condition on weakness of will attribution, 95% confidence interval for a direct effect = 0.68 to 2.10 (see Figure 2).

[9] Interestingly, again, participants in the Stereotypical condition more strongly agreed that *Jesse knew it was bad for him to continue drinking* than that *it was bad for him to continue drinking*, one-sample t-test, $t(30) = 2.45$, $p = .020$, MD = 0.515, test proportion = 5.13. Moreover, the same was true for participants in the Nonstereotypical condition with respect to continuing to add cards, $t(30) = 3.476$, $p < .002$, MD = 0.90, test proportion = 4.39.



*$p < .05$; **$p < .01$; ***$p < .001$

Fig. 2. Mediation results from Experiment 2. Left panel: Beth cover story. Right panel: Jesse cover story. Parenthetical values represent the strength of a simple regression between the two variables; values outside the parentheses represent the strength of the relationships in a model used to test for mediation.

3.3 Discussion

These results do not support the hypothesis that the stereotype effect is indirect and mediated by the implicit attribution of a commitment violated in the story. For neither cover story did the commitment attribution mediate the effect of condition. The results from Beth's story show that sometimes participants' own evaluation of the activity leads directly to the attribution of weakness of will and mediates the effect of stereotypes. But the results from Jesse's story show that such mediation does not always occur, suggesting that sometimes stereotypes directly affect weakness of will attributions. These results are consistent with the hypothesis that in stereotypical cases the absence of certain commitments can directly trigger attribution of weakness of will.

It's possible that we failed to observe a more robust mediation of the stereotype effect because we didn't look in quite the right place. Sousa and Mauro hypothesized that in cases of seriously immoral actions without explicit commitment violations, people attributed weakness of

will because they implicitly attributed a relevant commitment to the agent's "deep self," not just the agent. More recently, George Newman, Julian De Freitas, and Joshua Knobe directly tested the hypothesis that deep-self attributions explain the surprising effect of moral valence on weakness of will attributions noted by both May and Holton (2012) and Sousa and Mauro (forthcoming). More specifically, Newman and colleagues proposed that moral judgments impact deep-self attributions, which in turn impact weakness of will attributions. They found evidence that deep-self attributions partially mediate the effect of moral status on weakness of will attributions (Newman et al., forthcoming).

So perhaps a better candidate for a psychological explanation of our earlier findings is in terms of deep-self attributions. The hypothesis under consideration, then, is that the stereotype effect is indirect and mediated by deep-self attributions. The next experiment directly tests this hypothesis.

4. Experiment 3: Stereotypes and deep-self attributions

4.1. Method

Participants (N = 128)[10] were randomly assigned to one of two conditions, Stereotypical and Nonstereotypical. Each participant read a single story. We used the two Jesse stories from Experiment 2. Participants were asked to rate their agreement with the same weakness of will attribution as in Experiment 2. On a subsequent page, they were asked to rate their agreement with the following statement, modeled after the wording used by Newman and colleagues (forthcoming):

> By continuing to [add cards/drink], Jesse went against his true self — the person he truly is deep down.

[10] Thirty-eight female, aged 18-68, mean age = 30, 95% reporting native competence in English.

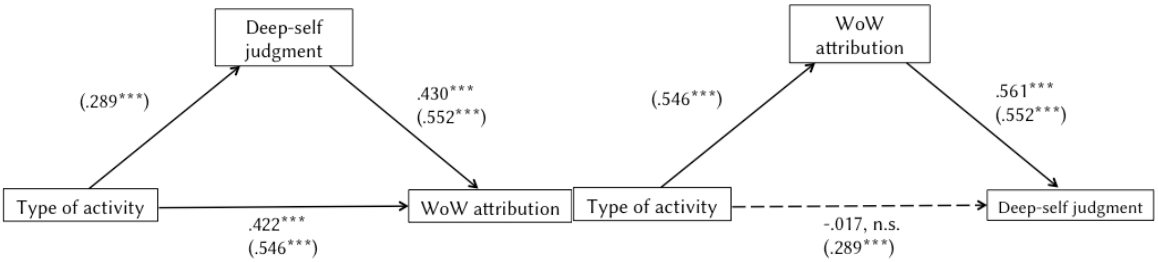
Responses were collected on a standard 7-point Likert scale, 1 (=Strongly disagree) to 7 (=Strongly agree).

4.2. Results

There was an effect of condition on weakness of will attribution, independent samples t-test, $t(126) = -7.322$, $p < .001$, MD = 2.05, d = 1.30, with scores in the Stereotypical condition (M = 5.03, SD = 1.56) higher than in the Nonstereotypical condition (M = 2.98, SD = 1.60). Attribution of weakness of will was significantly above midpoint (=4) in the Stereotypical condition, one sample t-test, $t(64) = 5.324$, $p < .001$, and below midpoint in the Nonstereotypical condition, $t(62) = -5.035$, $p < .001$. These results replicate the findings from the Jesse cover story from Experiment 2. There was also an effect of condition on deep-self judgments, independent samples t-test, $t(126) = -3.393$, $p < .001$, MD = 0.886, d = .60, with scores higher in the Stereotypical condition (M = 3.58, SD = 1.53) than in the Nonstereotypical condition (M = 2.70, SD = 1.42).

To test for the mediating role of deep-self judgments on weakness of will attributions, we conducted a bootstrap mediation analysis (Hayes 2013) with condition as the independent variable, agreement with the weakness of will statement as the dependent variable, and agreement with the deep-self statement as potential mediator. Deep-self judgments did *partially* mediate the effect of condition, 95% confidence interval for the indirect effect = 0.17 to 0.85. But the mediation was far from complete: the direct effect of condition was much larger than the indirect effect, 95% confidence interval for the direct effect = 1.08 to 2.09 (see Figure 3a).

(a) (b)



***$p < .001$

Fig. 3. Mediation results from Experiment 3. Parenthetical values represent the strength of a simple regression between the two variables; values outside the parentheses represent the strength of the relationships in a model used to test for mediation.

We also tested the reverse mediation model with condition as the independent variable, agreement with the deep self statement as the dependent variable, and agreement with the weakness of will statement as potential mediator. Weakness of will judgments fully mediated the effect of condition on deep self judgments: 95% confidence interval for the indirect effect = 0.56 to 1.44; 95% confidence interval for the direct effect = -0.59 to 0.49 (see Figure 3b).

4.3 Discussion

Newman et al. (forthcoming) found that deep-self judgments partially mediated the effect of moral valence on weakness of will attributions. Our results replicate their finding and, moreover, suggest that the mediating role of deep-self judgments extends beyond narrowly moral judgments to include broader normative considerations,[11] For our results are not plausibly due to a difference in moral valence. It might be imprudent to get drunk on an occasion, and a pattern of

[11] Knobe's previous work suggests that folk psychological attributions are sensitive to the valence of normative considerations that extend beyond the narrowly moral. See e.g. Pettit and Knobe (2009).

such behavior might amount to a moral failing due to, say, upsetting loved ones or burdening society with additional health-care costs in the long run. But getting drunk does not seem like a moral issue. And yet deep-self judgments still partially mediated weakness of will attributions, indicating that the mediating role of deep-self judgments transcends moral contexts.

The results from both mediation analyses suggest that we must exercise caution before inferring a specific causal relationship between judgments about weakness of will, on the one hand, and judgments about the deep self, on the other. For we observed mediation of condition in both directions. The fact that the one mediation was complete whereas the other was only partial suggests that if either judgment is causing the other, then weakness of will judgments are probably causing deep self judgments. But the present study was not designed to decide this question and further research is needed to settle the matter.

To sum up, the results from this experiment do not support the hypothesis that the stereotype effect is indirect and mediated by deep-self judgments. Rather, in this study the stereotype effect was *primarily direct* and only partly mediated by deep-self judgments. In other words, even controlling for the role played by deep-self judgments, there was a significant direct stereotype effect on weakness of will attributions.

These results are consistent with the hypothesis that in stereotypical cases the absence of certain commitments often directly triggers weakness of will attributions. Experiment 4 tests this hypothesis more directly.

5. Experiment 4: Stereotypes without practical commitments

5.1. Method

Participants (N = 61)[12] were tested. Each participant was asked to consider twelve short descriptions of someone who has never committed to a certain course of action. Six of the descriptions pertained to activities stereotypically associated with weakness of will (Stereotypical activities), and six were not thus associated (Nonstereotypical activities). For each group of activities, three featured acts of omission (i.e. to *avoid* certain courses of action) and three featured acts of commission (i.e. to *carry out* certain courses of action). Participants were asked to rate how weak-willed a person who satisfied each individual description was. Here are the instructions and twelve descriptions; the first six are Stereotypical, the final six are Nonstereotypical:

> Consider each of the following short descriptions. Suppose that a person fits the description. In virtue of fitting the description, how much weakness of will do you think the person exhibits?
>
> 1. Has never committed to avoid drinking alcohol
> 2. Has never committed to avoid recreational drug use
> 3. Has never committed to avoid cheating on his or her romantic partner
> 4. Has never committed to study for exams in school
> 5. Has never committed to watch his or her weight
> 6. Has never committed to exercise regularly
> 7. Has never committed to avoid international travel
> 8. Has never committed to avoid spicy food
> 9. Has never committed to avoid sky-diving
> 10. Has never committed to run a 10-mile race

[12] Twenty-eight female, aged 18–55, mean age = 30.8, 92% reporting native competence in English.

11. Has never committed to learn chess

12. Has never committed to go bungee jumping

Responses were collected on a standard 9-point Likert scale, 1 (=none at all) to 9 (=an excessive amount). The order of descriptions was rotated randomly.

5.2 Results

Responses to all the Stereotypical activities formed a reliable scale (Cronbach's $\alpha = .725$), as did the responses to all the Nonstereotypical activities (Cronbach's $\alpha = .840$). The mean score for all Stereotypical activities (M = 5.75, SD = 1.53) was significantly higher than the mean score for Nonstereotypical activities (M = 3.39, SD = 1.56), paired samples t-test, $t(60) = -8.30$, $p < .001$. The mean difference between Stereotypical and Nonstereotypical scores was 2.36 with a 95% confidence interval ranging from 2.93 to 1.79. The eta squared statistic was .534, which indicates a large effect size (Cohen 1988: 284-7). Mean response for the lowest Stereotypical activity (Alcohol, M = 5.12, SD = 2.39) was significantly higher than the mean response to the highest Nonstereotypical activity (10M race, M = 3.87, SD = 2.13), paired samples t-test, $t(59) = 3.309$, $p < .002$. (Table 1 includes the means for all the activities.)

Perhaps the most telling statistic is the mode response for the various activities. The mode for Stereotypical activities ranged from 6 to 9. By contrast, the mode for *each* Nonstereotypical activity was 1 (see Table 1).

	Stereotypical						Nonstereotypical					
	alcohol	**drug**	**cheat**	**study**	**weight**	**exercise**	**travel**	**spicy**	**Sky dive**	**10M-race**	**chess**	**bungee**
Mean	5.12	5.62	6.46	5.66	5.93	5.69	3.33	3.33	3.10	3.87	3.38	3.31
Mode	7	8	9	7	7	6	1	1	1	1	1	1

Table 1: Results from Experiment 4. 1= None at all, 9= An excessive amount.

5.3 Discussion

These results clearly support the hypothesis that, for at least many activities stereotypically associated with weakness of will, *lacking* practical commitments counts as being weak-willed. Thus it's probable that the ordinary notion of weakness of will includes at least one non-psychological factor that contributes significantly to weakness of will attributions: an agent can be weak-willed in virtue of failing to form relevant commitments regarding stereotypically weak-willed behavior.[13]

6. Experiment 5: Quality of Outcome

We now have evidence that the folk concept is sensitive to factors that are *upstream* of weak-willed episodes. In particular, failing to have formed a relevant commitment is often viewed as itself weak-willed. We also know, from previous work by both May and Holton (2012) and

[13] A referee suggests that weakness of will might not come in degrees, which if true would make responses in this experiment harder to interpret. In reply, we don't see any reason, either intuitive or theoretical, why weakness of will would be any different from strength of will or willpower in this respect, which both come in degrees. And while further empirical work could prove otherwise, we think it's perfectly intelligible to judge one person more weak-willed than another. In any event, the pattern in mode responses is still quite revealing.

Sousa and Mauro (forthcoming), that weakness of will attributions are also sensitive to the moral valence of the action itself. This naturally led us to wonder whether the folk concept is also sensitive to any *downstream* factors too: are weakness of will attributions sensitive to things that happen *after* the weak-willed action occurs? The most obvious candidate is *the quality of the action's outcome*. Perhaps agents are more readily judged to have displayed weakness of will when the outcome of their actions are *bad* rather than *good*.

6.1 Method

Participants (N=101)[14] were randomly assigned to one of four conditions in a 2 (Outcome: Good/Bad) x 2 (Commitment: Judgment/Resolution) between-subjects design: Good Resolution, Bad Resolution, Good Judgment, and Bad Judgment. We included the Commitment factor because prior theoretical and experimental work has identified two different types of practical commitment relevant to weakness of will: resolutions and normative judgments.[15] We did not expect an effect of Commitment and included it as a robustness check.

Each participant read a single story. Here is the basic story (with the Judgment/Resolution and Good/Bad outcome variations in brackets):

---

[14] Thirty-seven female, aged 18–77, mean age = 28, 98% reporting native competence in English. We excluded data from twenty-nine participants who failed comprehension questions. Including these participants in the analysis didn't significantly affect the results reported below.

[15] Other than the relevant mental state, the difference between the two conditions is that in Resolution (but not in Judgment) Peter does not violate a commitment that he has at the time of action, since he has *revised* his commitment. Such resolution-revision is, for Holton, characteristic of weakness of will. Previous experimental work (not reported here) revealed that participants overwhelmingly described *both* forms of commitment violations as instances of weakness of will.

Peter has been invited to a party on Thursday evening. The party will be a lot of fun. But he also has a test on Friday morning. It is important to Peter that he does well on Friday's test. He has concluded that if he goes to the party on Thursday evening, then he'll stay out late and do poorly on the test. So Peter [judges/firmly resolves] that he [definitely should not/will not] go to the party that evening. ¶ On Thursday evening, Peter's friends call to encourage him to go to the party. Peter reconsiders the situation. He agrees that the party will be fun, [but he still thinks that he definitely should not go to the party / so he explicitly abandons his earlier resolution and decides to go to the party]. [Nevertheless] He goes to the party [anyway] and stays out late. [Surprisingly/Unsurprisingly], he ends up getting a [100/30]% (an [A+/F]) on his test.

Participants then answered five comprehension questions, followed by the test question:

(WoW) By going to the party, does Peter display weakness of will?

Participants then rated how confident they were in their answer to the test question, from 1 (=not at all confident) to 10 (=completely confident).

6.2 Results

To test whether quality of outcome affects whether an action is view as weak-willed, we combined the two measures used to assess judgments about weakness of will: the dichotomous WoW choice and the confidence measure. We assigned a value of 1 to an answer of 'yes' to the WoW choice, and a value of -1 to 'no'. We then multiplied this value by the participant's confidence rating. Thus each participant's score for this calculated dependent measure fell on a

20-point scale ranging from –10 (maximum WoW denial) to +10 (maximum WoW ascription). We call this score a *weighted WoW ascription*.

Preliminary analysis revealed that, as expected, Commitment entered into no main or interaction effects, so for ease of exposition we collapse across that factor. An independent-samples t-test revealed an effect of Quality on weighted WoW ascription, $t(72.91) = -3.02$, $p = .004$, MD = 2.50, d = 0.71, with scores higher for Bad outcomes (M = 9.08, SD = 2.56) than for Good outcomes (M = 6.58, SD = 5.3).[16] Mean scores were well above the midpoint (=0) in both conditions, one-sample t-tests, ps < .001.

6.3 Discussion

These results are good initial evidence that quality of outcome affects whether an action is viewed as weak-willed. Previous work by both May and Holton (2012) and Sousa and Mauro (forthcoming) has shown that the quality of an outcome's moral valence can affect judgments about weakness of will. But performing well or poorly on a test is not plausibly viewed as a moral matter. So the moral *and* non-moral valence of an action's outcome can affect whether the action is viewed as weak-willed, which offers additional evidence that folk psychological attributions are sensitive to the valence of a broad range of normative considerations.

7. Conclusion

The philosophical debate about weakness of will has long assumed it to be a wholly psychological phenomenon, essentially involving the violation of a practical commitment.

[16] Follow-up studies (not reported here) revealed that when the outcome isn't specified at all (i.e. no mention of how Peter did on the test), participant response doesn't differ significantly from when a bad outcome is specified. This is unsurprising, since participants are likely to infer a bad outcome, given the details of the case.

Recent debates have focused on which type of violation is essential to the ordinary notion of weakness of will. However, when it comes to the ordinary notion, we've provided strong evidence that this assumption is mistaken.

While the ordinary concept of weakness of will is unquestionably sensitive to the violation of practical commitments Experiment 1 revealed that a surprising number of people strongly attributed weakness of will in the absence of such violations, which suggests that the ordinary concept of weakness of will might include important non-psychological components. We found evidence that the ordinary concept is indeed sensitive to non-psychological factors. In particular, we found evidence that, at least for conduct stereotypically associated with weakness of will, the *failure to form* such commitments directly triggers weakness of will attributions (Experiments 2-4). Furthermore, we found evidence that the ordinary concept is sensitive to a second non-psychological factor, namely, the quality of the outcome (Experiment 5).

When assessing whether someone's action was weak-willed, people attend to more than just whether the action features a failure by the agent to abide by a practical commitment. While such failures are certainly part of the ordinary notion of weakness of will, so too are considerations that lie both downstream and upstream from such violations. Downstream, we have shown that people attend to the quality of the outcome of actions, with bad outcomes contributing to more emphatic weakness of will attributions. Even more tellingly, we have shown that upstream from any commitment-violating actions, for actions stereotypically associated with weakness of will, a failure to form relevant commitments can trigger weakness of will attributions. In showing that the folk concept includes such non-psychological elements, we have shown that it is much broader than the philosophical debate has thus far imagined.

What are the consequences for philosophical theorizing about weakness of will of the discovery that the folk concept involves the failure to form commitments in stereotypical cases?

We can imagine at least three possibilities. The first is that our findings show that the folk concept of weakness of will includes a range of disparate and unrelated elements, and that philosophers have been mistaken in attempting to give a unified account of the phenomenon. The second is that the folk are simply mistaken in identifying the failure to form commitments as weakness of will, perhaps because the term is a piece of technical vocabulary or because the nature of survey-based experiments does not allow them to clearly distinguish the various forms of failures in this area.

But there is a third possibility, one that unifies our findings with previous philosophical theorizing. It is that the folk concept of weakness of will involves the failure to exercise appropriate self-control. Narrowly conceived, such self-control involves the ability to do what one has committed to doing: to keep one's resolutions, act on one's judgments, and resist temptation and the tendency to rationalize. This narrow version of self-control is captured by the traditional, purely psychological accounts of weakness of will. More broadly, however, self-control involves a higher-level skill in forming commitments that are conducive to achieving one's ultimate ends. On this view, one fails to exercise proper self-control by failing to recognize the ways in which one's pursuit of those ends can be thwarted by a range of temptations and distractions, and so by failing to form the kinds of commitments that would help one to avoid such temptations and distractions. On such a model, never committing to e.g. avoiding recreational drugs represents a failure of self-control because such a commitment would help to forestall behavior that harms one's pursuit of one's ultimate ends. In failing to form the commitments that would best serve the pursuit of such ends, one fails to properly control one's desires, not because one succumbs to them despite committing not to, but because one never

realizes that such commitments are necessary.[17] This would explain why it is particularly in stereotypical cases that the lack of practical commitment earns weakness of will attributions, since such cases involve the kinds of temptations that are taken to represent barriers to the successful pursuit of commonly assumed shared ends, such as health, wealth, and happiness.

This model, then, treats weakness of will as a failure of practical reason that can have an evaluative as well as a volitional element. The standard view of weakness of will has treated it as the failure of a distinct *volitional* capacity — a failure to form or carry out a commitment to do what one has judged to be best.[18] This has standardly been treated as distinct from the *evaluative* capacity to form good judgments about what one has most reason to do. Our view, however, is that weakness of will, understood broadly as a failure of self-control, can also involve a failure that lies midway between the evaluative and the volitional. It is a failure to form commitments that are properly required by one's long-term goals, and so it is volitional. But it is not a failure to form commitments that reflect evaluative judgments about what it would be best to do: rather, it involves a failure to *recognize* what it would be best to do in light of one's overall goals. In this sense, the failure is evaluative.

The view that weakness of will can be displayed by simply failing to form particular rational commitments means that our account joins others in including an important normative element. For example, Holton distinguishes weak-willed resolution revision from mere changes of mind in light of an account of what makes such revisions rational (2009). Beebe has

---

[17] Jeanette Kennett (2014) has recently defended a model of addiction as lack of self-control that includes a similar suggestion. According to Kennett, some addicts "simply cannot conceive of the value to be found in dedication to long-term projects" (p. 157), while others can conceive of such value but do not believe that such lives are open to them. Both of them therefore fail to form the sorts of commitments necessary to secure such values. Kennett counts both forms of addiction as failures of self-control. We thank a reviewer for *Synthese* for bringing Kennett's paper to our attention.

[18] In addition to Mele (2012) and Holton (2009), a similar view is defended in e.g. Smith (2003) and Cohen & Handfield (2010).

confirmed that a version of this distinction is part of the folk conception of weakness of will, which is sensitive to the rationality of the both the revised resolution and the subsequent action (2013). May & Holton (2012) and Sousa & Mauro (forthcoming) also show that the folk conception is sensitive to the moral valence of the action. What sets our proposal apart from these normative accounts is that the normativity is well upstream of both resolution and action: in our view, it is possible to be weak-willed by failing to form commitments that you ought to form, and not merely by irrationally revising such commitments or by acting in ways that violate other normative standards.

Our proposed model therefore accounts for the tendency to attribute weakness of will in stereotype cases that lack commitment violations. It can also partially account for the influence of outcomes on weakness of will attributions. Some cases in which agents do not attribute weakness of will to violations that lead to good outcomes will be simple performance errors, in which the outcome distracts from the actual weakness of will. In other cases, however, an unwillingness to attribute weakness of will to good outcomes is precisely what the model would predict. We've suggested that self-control involves the exercise of skill in forming and managing commitments in order to achieve one's long-term goals. Bad outcomes are more clearly inconsistent with achieving those goals than good outcomes. In fact, good outcomes that follow commitment violations can suggest that the commitment was not necessary in order to secure the long-term goal, and so that abandoning it was consistent with the proper exercise of the skill of self-control.

On Holton's account of weakness of will, revising a resolution can be rational, rather than weak-willed (2009), something that Beebe has shown fits with folk ascriptions of weakness of will (2013). Holton's account does not define rational revisions as those that bring about good outcomes. Rather, he claims that rational revisions are those that manifest a tendency to revise

resolutions that it would be rational to have. While this means that many revisions that happen to lead to good outcomes will still count as irrationally weak-willed, we can nevertheless understand such a tendency partly in terms of the disposition to produce good outcomes. So the decreased confidence that commitment violation is weak-willed in good outcome cases might join the tendency to attribute weakness of will in stereotypical cases as evidence that such attributions are sensitive to considerations about which commitments are necessary to secure long-term self-control, as well as to more explicit commitment violations.

This account of self-control therefore involves two distinct conditions. The first is a matter of the *structure* of an agent's will — the relationship between her actions, her judgments, and her intentions. To be self-controlled is to have these in proper alignment, while to be weak-willed is for them to conflict in certain ways. This first condition has dominated the debates on weakness of will. Much less discussed has been the importance of the *content* of her will: the particular commitments that she has formed. A will lacking in certain specific judgments, intentions, resolutions, or dispositions can be weak despite having a structure free from internal conflicts. Self-control therefore involves both the relationship between the parts of an agent's will, and the identity of the parts themselves. We can be weak-willed either by putting the parts together in the wrong way, or by simply not including the right parts in the first place. To be properly self-controlled, an agent should not only avoid internal conflict between commitment and action: she should also *form* a range of commitments and *avoid* forming others. On this unified two-tier account, then, self-control involves more than just strength of will in resisting temptation. It also involves the exercise of practical wisdom in setting ends and forming commitments.

Acknowledgements

For helpful discussion and feedback, the authors thank Wesley Buckwalter, Joshua Knobe, Alfred Mele, Josh May, James Beebe, Ryan Ehrlich, Richard Forbes, Natalie Galloway, Julia Hill, Charles Millar, Jay Solanski, Ayomide Yomi-Odedeyi, Angelo Turri, Christine Tappolet, the audience at the 2014 Canadian Philosophical Association meetings at Brock University, and the anonymous referees at *Synthese*. This research was supported by the Social Sciences and Humanities Research Council of Canada and an Early Researcher Award from the Ontario Ministry of Economic Development and Innovation.

Works Cited

Aristotle. *Nicomachean Ethics*. (1999) T. Irwin (trans.) Indianapolis: Hackett Publishing.

Beebe, J. (2013). Weakness of will, reasonability, and compulsion. *Synthese,* 190: 4077-4093.

Beebe, J. (Under review). The folk conception of weakness of will.

Cohen, J. W. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, D and Handfield, T. (2010). Rational Capacities, Resolve, and Weakness of Will. *Mind,* (119): 908-932.

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge: Cambridge University Press.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.

Holton, R. (1999). Intention and Weakness of Will. *Journal of Philosophy* (96): 241–262.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford: Clarendon Press.

Kennett, Jenette. (2014). Just Say No? Addiction and the elements of self-control. In Levy, N.,

ed. *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*. Oxford University Press: 144-164.

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, 102(1), 72. doi:10.1037/0033-2909.102.1.72

May, J. and Holton, R. (2012). What in the world is weakness of will? *Philosophical Studies* (157): 341-360.

Mele, A. (1987). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. Oxford University Press.

Mele, A. (2010). Weakness of Will and Akrasia. *Philosophical Studies* (150): 391-404.

Mele, A. (2012). *Backsliding: Understanding Weakness of Will*. Oxford University Press.

Newman, G., Knobe, J., & Bloom, P. (2014). The moral nature of the true self. *Personality and social psychology bulletin* (40): 203-216.

Newman, G., De Freitas, J., & Knobe, J. (Forthcoming). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science.*

Johnson, J. T., Robinson, M. D., & Mitchell, E. B. (2004). Inferences about the authentic self: when do actions say more than mental states? *Journal of personality and social psychology*, 87(5), 615.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3), 279–301.

Smith, M. (2003). Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion. In Stroud, S and Tappolet, C. eds. *Weakness of Will and Practical Irrationality*. Clarendon Press: 17-38.

Sousa, P and Mauro, C. (forthcoming). The Evaluative Nature of the Folk Concepts of Weakness and Strength of Will. *Philosophical Psychology*.