Is a Bad Will a Weak Will?

Cognitive Dispositions Modulate Folk Attributions of Weakness of Will

Forthcoming in *Philosophical Explorations*

Alejandro Rosas
Universidad Nacional de Colombia
arosasl@unal.edu.co

Juan Pablo Bermúdez
Universidad Externado de Colombia
juan.bermudez@uexternado.edu.co

Jesús Antonio Gutiérrez Cabrera
Universidad Nacional de Colombia
jeagutierrezca@unal.edu.co

Abstract

In line with recent efforts to empirically study the folk concept of weakness of will, we examine two issues in this paper: (1) How is weakness of will attribution [WWA] influenced by an agent's violations of best judgment and/or resolution, and by the moral valence of the agent's action? (2) Do any of these influences depend on the cognitive dispositions of the judging individual? We implemented a factorial 2x2x2 between–subjects design with judgment violation, resolution violation, and action valence as independent variables, and measured participants' cognitive dispositions using Frederick's Cognitive Reflection Test [CRT]. We conclude that intuitive and reflective individuals have two different concepts of weakness of will. The study supports this claim by showing that: a) the WWA of intuitive subjects is influenced by the action's (and probably also the commitment's) moral valence, while the WWA of reflective subjects is not; b) judgment violation plays a small role in the WWA of intuitive subjects, while reflective subjects treat resolution violation as the only relevant trait. Data were collected among students at two different universities. All subjects (N=710) answered the CRT. A three-way ANOVA was first conducted on the whole sample and then on the intuitive and reflective groups separately. This study suggests that differences in cognitive dispositions can significantly impact the folk understanding of philosophical concepts, and thus suggests that analysis of folk concepts should take cognitive dispositions into account.

Keywords: Weakness of will; cognitive dispositions; experimental philosophy; folk concepts; reflection; intuition.

1. Introduction

Recent empirical approaches to weakness of will attribution

When we say that someone is weak-willed, what do we mean? The two main philosophical accounts portray weakness of will either as a failure to stick to one's best, all-things-considered judgment (Mele 1995) or as a failure to abide by one's resolution in the face of temptation (Holton 1999). After gathering some data about folk uses of the concept, Mele and Holton have arrived at an agreement that both judgment violation and resolution violation seem to be a part of the concept, given the way people tend to attribute weakness of will (Mele 2010; May & Holton 2012). But questions remain about the relative weights of judgment violation and resolution violation, and about the influence of the moral valence of the action on weakness of will attribution [WWA]; and more generally, about whether these influences systematically depend on personality dimensions like intuitive or reflective cognitive dispositions.

Mele (2010) suggested that judgment violation and resolution violation are independently sufficient for WWA, i.e., that the two criteria work *disjunctively*. He tried to produce an empirical proof of the sufficiency of judgment violation by registering participant responses to a vignette about an agent who acts against his better judgment but explicitly avoids forming any resolution. He found that a majority of subjects attribute weakness of will in that case. May & Holton (2012) pointed out that the vignette allows an interpretation of the agent as forming a second-order resolution (to make a first-order resolution at an appropriate time) that he then violates. To more comprehensively address the issue, they developed a factorial 2x2 design with 4 vignettes where the agent violated her resolution, her best judgment, both, or neither. This would allow them to check for independent main effects of resolution violation and judgment violation, as well as their interaction. They carried out two experiments. The first experiment's vignettes lacked uniformity (the agent's actions had various valences, two representing adultery with different motivations, another one a courageous but imprudent action, and yet another one the violation of a diet). Since uncontrolled variables may have interfered, they ran a second experiment with uniform vignettes in which an agent either performs or does not perform a skydiving jump violating

either his prior judgment, his resolution, both or neither. Both experiments yielded roughly similar results: a significant main effect for each violation independently, but significantly higher scores when both violations occurred together. This suggested that weakness of will is not a disjunctive concept, but rather works like a *cluster* of both violation criteria. In May and Holton's second experiment, the courageous nature of the action (skydiving) seems to have pulled WWA down to the midpoint when the agent jumps against both his best judgment and resolution. Therefore, apparently, the action's positive valence had a significant influence in some participants' WWA independently of commitment violations. In a third experiment, May and Holton investigated whether the action's valence influenced WWA, and confirmed that the action's immoral valence had a main effect on WWA independently of whether the agent violates or conforms to her prior commitments.

Two subsequent papers bracketed the issue of the relative weights of judgment and resolution violation, and focused on the influence of the action's valence on WWA. Sousa & Mauro (2014) merged judgment and resolution into a single commitment that was either moral or immoral, and paired it with either a moral or an immoral action, thereby creating four vignettes: 1) moral commitment and immoral action; 2) moral commitment and moral action (these two they called *'prototypical'*, presumably since they are the situations commonly associated with the concepts of weakness and strength of will, respectively); 3) immoral commitment and immoral action; and 4) immoral commitment and moral action (these two they called *'non-prototypical'*, since they are less commonly associated with the traditional concepts). Sousa & Mauro (2014) predicted that ordinary people would hesitate to attribute strength of will in 3) and weakness of will in 4), even though this is what the philosophical concepts would dictate. The reason for hesitation is the existence of an evaluative tension between the action's immorality (negative) and the agent's compliance with his prior commitment (positive) in 3), and between the moral action (positive) and the agent's violation of his prior commitment (negative) in 4). But they found that in 3) and 4) (i.e. the non-prototypical cases) responses formed a threefold pattern: a) some participants hesitated to attribute weakness or strength of will, probably due to the evaluative tension mentioned above; b) other participants gave orthodox responses, attributing weakness of will only when the agent violated her commitment, irrespective of the valence of actions or

commitments; and c) another group of participants gave unorthodox responses, attributing weakness of will when the agent acted immorally, and strength of will when he acted morally (probably guided by the action's valence rather than the commitment violation or compliance). To explain the latter group's unorthodox responses – which radically departed from the philosophical concept –, Sousa & Mauro postulated that those participants were implicitly attributing judgments and/or resolutions to the agent: thus, when the agent acted immorally, participants attributed an implicit moral commitment whose violation resulted in weakness of will; and when the agent acted morally, they attributed an implicit moral commitment whose adherence resulted in strength of will. Thus, their reading transformed unorthodox responses into orthodox ones, against the literal reading of the vignettes. (Here it is important to point out that this hypothesis can find support in the wording of their vignettes: the agents are in fact described as being in conflict between moral and immoral commitments, or at least as pondering both alternatives before deciding. Thus, the vignettes themselves would give participants some ground to attribute implicit commitments to the agent.)

Doucet & Turri (2014) questioned this explanation. They reported evidence that the action's valence influences WWA directly, without the mediation of implicit commitment attributions. Their study 2 tested whether WWA was mediated by implicit attributions of moral commitments, and yielded negative results. In contrast to Sousa & Mauro (2014), Doucet and Turri's vignettes do not describe agents as being in conflict before forming their commitment. Participants thus had no encouragement to make implicit attributions.

What then accounts for the unorthodox responses of participants who depart from the philosophical weakness-of-will concept? If no implicit commitment attribution mediates such responses, then moral valence seems to directly influence WWA. But why then would the influence of moral valence be stronger in some participants than in others? An interesting hint may be given by Sousa & Mauro, who suggested that participants whose the unorthodox responses departed from the philosophical concept would tend to have an *intuitive* rather than *reflective* cognitive disposition (Frederick 2005). Doucet and Turri did not control for the cognitive dispositions of participants, so this hint remains empirically

untested. It would mean that moral valence directly influences WWA (as Doucet and Turri contend), but that this influence depends on having an intuitive cognitive disposition. The rationale behind this is that immorality and weakness of will tend to be strongly intuitively associated, since they are both negative evaluative traits. To explain this further we introduce the notion of cognitive dispositions.

Cognitive dispositions and weakness of will attribution

When you face a question and intuitively come up with an answer, what do you do? Do you endorse your intuitive answer right away, or do you examine the intuition critically before endorsing it? We call people who tend to do the former *intuitive*, and those who do the latter *reflective*. These are the two cognitive dispositions that are relevant for the present study. The distinction between intuitive and reflective dispositions relies on a broadly dual-process account of human cognition, which distinguishes between *automatic* or *intuitive* processes that can be performed independently from working memory, and *reflective* processes whose performance requires the use of working memory (Evans 2010; Evans & Stanovich 2013). Here 'working memory' (Baddeley 2007) refers to the set of higher-order cognitive capacities that allow for the mental manipulation of task-oriented representations. Intuitive processes tend to be fast, effortless, and require no concentration; reflective processes tend to be slower and require cognitive effort and concentration. Intuitive people tend to rely on heuristic, rough-and-ready responses to questions and problems, while reflective people tend to double-check those heuristic solutions by means of step-by-step, working-memory-reliant processes.

Cognitive dispositions are personality traits concerning how likely a person is to engage in reflective processes to assess the outcomes of her own intuitive processes. This can be measured by means of the Cognitive Reflection Test (or CRT), whose questions are designed to elicit intuitive answers that are nonetheless incorrect. In order to reach the right answer, the participant would have to reflectively inhibit the intuitive response and concentrate on seeking the right answer (Frederick 2005). The CRT consists of three questions, and the participant score is the number of questions she gets right.

Consider again the issue at hand. Studies of folk WWA have found that some participants make unorthodox weakness of will attributions: they attribute weakness of will to an agent who sticks to an immoral commitment by performing an immoral action; and they attribute strength of will to an agent who breaks an immoral commitment by performing a moral action (Sousa & Mauro 2014). How can we explain this if, as Doucet and Turri's (2014) work suggests, participants do not seem to implicitly attribute a commitment to the agent?

The answer may lie in the difference in cognitive dispositions. An immoral action would tend to intuitively and automatically elicit high WWAs, regardless of considerations of violation or conformity with prior commitments. Intuitive participants would tend to endorse the initial, intuitive high WWA. Only people with a reflective disposition would be likely to inhibit the automatic, intuitive tendency to group immoral action together with weakness of will, and moral action together with strength of will, and focus instead on violation or conformity with prior commitments.

Our aim in this paper is to investigate the influence of three variables on folk attributions of weakness of will: judgment violation, resolution violation, and moral valence of action. We also take into account the difference between intuitive and reflective dispositions of participants. To this end we implement a 2x2x2 between–subjects design and use the Cognitive Reflection Test [CRT] to measure participants' cognitive dispositions. This design is intended, first, to test whether the concept of weakness of will works disjunctively or as a cluster by measuring judgment violation, resolution violation and action valence simultaneously; and, second, to test Sousa & Mauro's prediction that intuitive and reflective participants will display different WWA patterns.

2. Materials and Procedures

Data were collected from students attending different courses at two universities. In total 710 students participated, mean age = 21 yrs., 39% female. Four were excluded for failing to complete the questionnaire. We devised 8 vignettes corresponding to the 2x2x2 design, where the only changes in the story concern the best judgment, the resolution made and the action performed by the protagonist. The story was kept as homogeneous as possible

across all conditions (avoiding the use of different stories and characters for different conditions), and as simple as possible (eliminating explicit suggestions of hesitation or inner conflict), to avoid confounding factors. The story was presented in Spanish, the participants' native language. Here is an English translation (see Appendix for original version):

> Rodrigo and Alenka have been friends since childhood, and they live in the same house. Rodrigo is a student and Alenka has a job. Rodrigo has just won a considerable sum of money in a contest, but has not yet told Alenka about it. Unfortunately, she has just learned that she has contracted a deadly disease, whose only effective treatment is very expensive. Her savings are not enough. She tells Rodrigo about this, and he realizes that the sum he has just won would cover the remaining costs. He also knows that Alenka has no living relatives and that he will inherit her savings if she dies. He must decide in 24 hours and notify the hospital, because the treatment must begin immediately and be paid in advance.
>
> After pondering the alternatives, he concludes that the best thing to do is to (wait for Alenka to die and inherit her savings / help Alenka). (Therefore / However), he resolves to (await her death/ help her). (Accordingly / However), when the time comes to notify the hospital, (he remains silent about his money and sits tight waiting for Alenka to die in order to inherit her savings / he announces that he will help her, and uses his contest money to pay for the treatment in advance.)

The 8 stories differ only in whether the protagonist judges it best to spend his money (won in a contest) to save the life of a close friend by covering the cost of her medical treatment, or to keep the money for himself; whether he resolves to do the former or the latter; and whether he finally spends his money on his friend's medical treatment or not. All other elements in the story remain the same. We thus have eight different conditions:

(1) IM¬V = Immoral action and no violation of commitments
(2) IM.JV = Immoral action and violation only of judgment
(3) IM.RV = Immoral action and violation only of resolution
(4) IM.JV&RV = Immoral action with judgment violation and resolution violation
(5) MO¬V = Moral action and no violation of commitments
(6) MO.JV = Moral action and violation only of judgment
(7) MO.RV = Moral action and violation only of resolution
(8) MO.JV&RV = Moral action with judgment violation and resolution violation

Particularly, these stories include what Sousa & Mauro (2014) have called prototypical and non-prototypical conditions. Their 2 non-prototypical conditions (immoral commitments followed by immoral action ["IR & IA"], and immoral commitments violated by a moral action ["IR & MA"]) correspond to conditions (1) and (8) respectively, and their 2 prototypical conditions (moral commitments followed by moral action ["MR & MA"], and moral commitments violated by immoral action ["MR & IA"]) correspond to conditions (5) and (4) respectively. We additionally have 4 'mixed' conditions in which the agent violates his judgment or his resolution, but not both, while acting immorally (2–3) or morally (6–7). These 'mixed' conditions allow us to measure the independent significance of each variable: judgment violation, resolution violation, and action valence.

Responses were collected with pen and paper in a between-subjects design: each participant was shown only one of the 8 vignettes. Participants were asked to state to what extent they agreed with the statement "In this story, Rodrigo shows weakness of will." Their answers were recorded in a Likert scale ranging from 0 (100% disagreement) to 10 (100% agreement), with 5 as a neutral midpoint. Additionally, all participants completed the CRT. The order of presentation between CRT and vignette was counterbalanced for all conditions.

In his seminal paper, Frederick (2005) formed an "intuitive group" consisting of all participants who had had no correct answers in the CRT, and a "reflective group" consisting of all participants who had given correct answers for all 3 questions. He thus excluded

participants with only 1 or 2 correct answers from his analysis. We departed slightly from Frederick's procedure in constructing our intuitive and reflective groups: our resulting reflective group would have been too small for the ANOVA, so we included in it participants with at least 2 correct answers.

3. Results

A three way ANOVA on weakness of will attributions in the whole sample (N=513 after excluding 193 participants who answered only 1 of the CRT questions correctly) showed a significant main effect for both judgment violation (*F (1, 505) = 9.58, p = .002,* η_p^2 *= .019*) and resolution violation (*F (1, 505) = 20.99, p = .000,* η_p^2 *= .040*). No interaction effects were observed. Subsequent three-way ANOVAs were performed separately on the intuitive and reflective groups. In the intuitive group (N=286), we found a significant main effect only for judgment violation (*F (1,278) = 6.59, p = .011,* η_p^2 *= .023*), and a close to significant main effect for resolution violation (*F (1,278) = 3.772, p = .053,* η_p^2 *= .01*). In the reflective group (N=227), we found a significant main effect only for resolution violation (*F (1, 219) = 20.95, p = .000,* η_p^2 *= .087*), but no significant main effect for judgment violation (*F (1,219) = 2.660, p = .104,* η_p^2 *= .012*). Action valence was close to significance in the intuitive group (*F (1, 278) = 3.38, p = .067,* η_p^2 *= .012*), but quite far from being significant for the reflective group (*F (1, 219) = .073, p = .79,* η_p^2 *= .000*). In the intuitive group there was a statistically significant interaction between resolution and judgment (*F (1,278) = 5.68, p = .018,* η_p^2 *= .020*), and between resolution and action valence (*F (1, 278)= 4.17, p = .042,* η_p^2 *= .015*). No interaction effects were found in the reflective group.

4. Discussion

Our results for the whole sample partially replicate the results of May & Holton's (2012) experiment 2. Violations of both judgment and resolution had a significant main effect on weakness of will attributions. Our *F* and *p* values were similar to theirs (May & Holton 2012, fn. 8). No interaction effects were observed. We did not get clear-cut evidence for or against the disjunctive or the cluster interpretations of the concept in the whole sample. A pairwise comparison of means by vignettes revealed that when the action is immoral (conditions 1–

4), only the violations of judgment and resolution together (condition 4) were significantly different from violation of neither (condition 1) ($p = .007$). This supports May & Holton's cluster interpretation; but when the action is moral (conditions 5–8), each commitment violation, and the violation of both together (conditions 6–8), were significantly different from no violation (condition 5), supporting Mele's (2010) disjunctive interpretation (all $p < .005$; see Fig. 1). Thus, for the group as a whole, it seems that weakness of will works as a *disjunctive* concept when the agent under consideration performs a *moral* action (so that the violation of any commitment can sufficiently motivate the attribution of weakness of will), but it works as a *cluster* concept when the agent performs an *immoral* action (so that only the violation of both commitments together makes the attribution significantly different from the violation of neither).

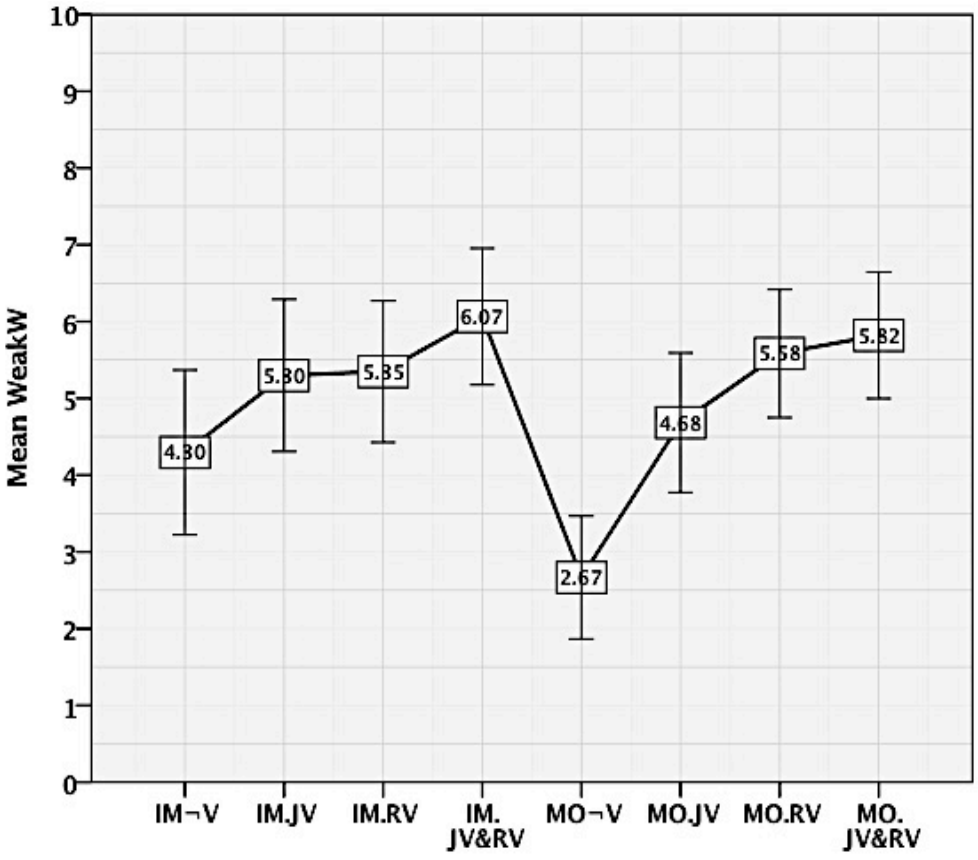

Figure 1: Mean weakness of will by vignette in the whole sample. Mean scores are represented on a scale from 0 to 10. Error bars: 95% CI

However, when analyses are performed on the intuitive and reflective groups separately, the results are drastically different: in the intuitive group only judgment violation had an effect on weakness of will attribution, whereas in the reflective group only resolution violation had an effect (see all mean WWA by group and condition in table 1 and table 2 in the appendix). In the intuitive group we observed two interaction effects: between resolution and judgment, and between resolution and action valence. A pairwise comparison of means by vignettes revealed that the effect of both judgment and resolution was present only in the moral action conditions (MO), but absent in the immoral action conditions (IM). In the moral action cases, each violation separately and both together were significantly different from violation of neither ($p = .000$ in all cases with moral action; $p > .1$ in all cases with immoral action); in the immoral action cases, the mean scores of weakness of will attribution ranged between *4.94* (lowest) when neither commitment was violated, and *6.05* (highest) when judgment was violated (5 being the neutral score), with no significant difference between them ($p = .191$). Thus, since neither violation of, nor compliance with, either judgment or resolution made a significant difference to WWA in cases of immoral action, it must be the action's immoral valence that predominantly guides WWA for intuitive participants in conditions 1–4. In spite of this, the action's valence had no main significant effect in the intuitive group (though it did come close to significance, $p = .067$). Statistically, this is due to the fact that the action's moral valence did not drive WWA in the case of moral actions violating immoral commitments (conditions 6–8). Conceptually, this is certainly an intriguing result demanding a careful explanation.

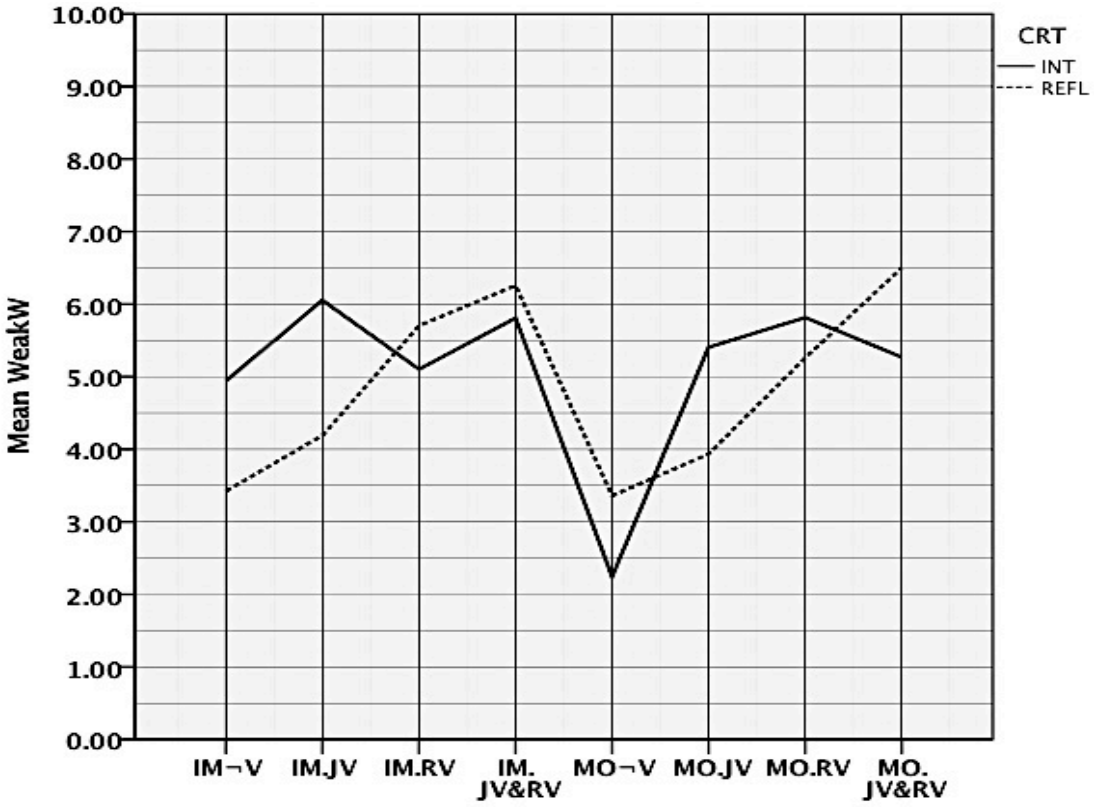

Figure 2: Mean weakness of will attributions by vignette in the reflective (dotted line) and intuitive (full line) groups. Mean scores are represented on a scale from 0 to 10.

The result is intriguing in light of a very plausible hypothesis by Sousa & Mauro (2014). In their experiment 2, the attributions of most participants in the non-prototypical conditions – conditions (1) and (8) in the present study – contradicted the philosophical concept of weakness of will: *63%* denied strength of will when the agent complied with his immoral commitments acting immorally, and *60%* denied weakness of will when the agent violated his immoral commitments acting morally. Sousa & Mauro (2014, 502) hypothesized that these participants probably had an intuitive cognitive disposition, whereas those who made attributions consistent with the traditional philosophical concept in these conditions probably had a reflective disposition.

We confirmed their hypothesis with respect to the reflective group: reflective participants strictly apply the philosophical concept in the two non-prototypical conditions, attributing less weakness of will in condition (1) (immoral action with no violation of immoral commitments), and greater weakness of will in condition (8) (moral action with violation of

immoral commitments). Moreover, the attributions of reflective participants were otherwise unaffected by the action's moral valence: a pairwise comparison of means reveals no significant difference in WWA between condition (1), where the agent acted immorally consistently with his immoral commitments, and condition (5), where the agent acted morally consistently with his moral commitments ($p = .950$). If moral valence had had a significant effect on WWA, attribution levels would have been significantly greater in case (1) than in case (5). Thus, Sousa and Mauro seem to be right about the reflective participants: a reflective disposition is correlated with moral valence having no influence on WWA.

But our data only partially corroborate their hypothesis regarding the intuitive group. The hypothesis predicted that the valence of the action would drive WWA. That is, when the action is immoral, it predicts that there would be no statistically significant difference in WWA scores between condition (1), where no violation occurs, and conditions (2-4), where either or both violations occur, simply because of the immoral valence of the action; and when the action is moral, it predicts that there would be no significant difference in WWA scores between conditions (6-8), where violation occurs, and condition (5), where no violation occurs. But only the former was confirmed. The mean for intuitive participants in condition (1) was *4.94*, relatively high for a condition where neither commitment is violated, and not significantly different ($p = .191$) from the scores of conditions (2-4) where the agent violates either or both commitments (the highest mean score for the intuitive participants in these conditions was *6.05* in condition (2)). So far, Sousa and Mauro's hypothesis succeeds in predicting the WWA of intuitive participants. But their scores in conditions (6-8) are significantly different from the score in condition (5) (see Fig. 2, continuous line). Our intuitive participants do not distinguish (statistically) between agents who act immorally while abiding by immoral commitments (condition 1), and agents who act immorally while violating moral ones (conditions 2-4) But they do distinguish between agents who act morally while abiding by moral commitments (5) and those who act morally while violating immoral commitments (6-8) ($p= .001$) (see Fig 2, continuous line).

We must therefore conclude that, although there is a difference in WWA between the intuitive and reflective groups, it is not fully explained by Sousa and Mauro's hypothesis: it is not the case that intuitive participants were fully guided by the action's moral valence in all non-prototypical conditions. Particularly, when the agent acted morally violating prior commitments (conditions 6–8), the action's moral valence did not lead them to attributing low levels of weakness of will.

Focusing now on condition (8), which corresponds to Sousa & Mauro's (2014) non-prototypical case where the agent violated immoral commitments by acting morally ("IR & MA"), our intuitive participants were in overall disagreement (see Fig. 3). In this condition, 46% of participants (labeled "-1", N=17) attributed scores of weakness of will higher than 5; 35% (labeled "1", N =13) attributed scores lower than 5; and 19% (labeled "0", N=7) gave a neutral attribution of 5. According to Sousa & Mauro's hypothesis, a majority should have given WWA scores lower than 5 and statistically equivalent to that of condition (5), where the action is moral and no commitments are violated.

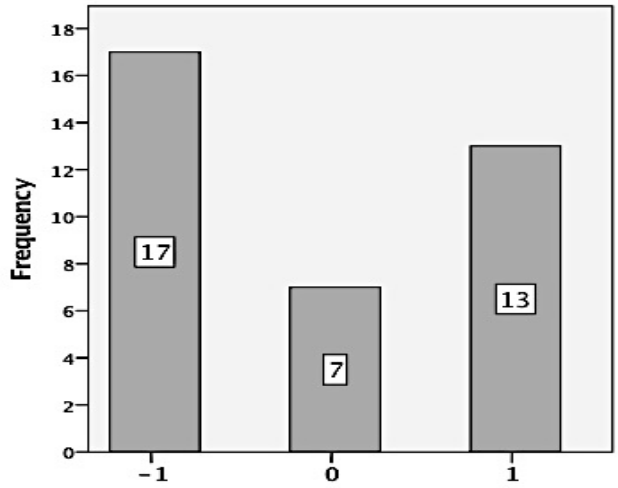

Figure 3: Distribution of responses for the intuitive participants in the non-prototypical case where the agent acts morally violating immoral judgment and immoral resolution. Weakness of will = -1; No weakness of will = 1; Neutral = 0.

Thus it is no wonder that action valence failed to reach the significance level of *.05* in the intuitive group (though it almost did): action valence explains WWA in all the conditions

with *immoral action*, but not in all the conditions with moral action. Why were intuitive participants apparently insensitive to the action's valence in conditions 6-8? A possible explanation lies in wording differences between our vignettes and those of Sousa & Mauro. In all their vignettes, the agent is portrayed as undergoing an inner conflict between moral and immoral motivations: he "seriously envisages" the immoral action, but is also "strongly inclined" to act morally. Had we formulated our vignettes in the same way, perhaps participants would have interpreted the agent in conditions 6–8 as being swayed by his previous moral considerations and would have given low WWA scores. However, this explanation cannot be swiftly applied to our findings, because, if the wording caused problems in the moral conditions, it should have caused the same effect in the immoral conditions, and it obviously did not.

Thus, it remains to be explained why the moral valence of the action did not decrease the intuitive participants' WWA score in conditions (6-8).

We see here two possibilities, although we shall not develop them in full. The first one has to do with difficulties in the CRT as a test of reflective dispositions. As some have argued (Sinayev and Peters 2016), the CRT measures numeric abilities rather than just dispositions to inhibit intuitive responses. Participants may have failed the CRT for lack of numerical abilities and thus, though reflective, were classified as intuitive, but nonetheless focused on the violation of commitments in the case of moral actions.

More likely, however, the relevant explanation for this pattern is that intuitive participants' susceptibility to moral valence extends beyond actions, and also includes the moral valence of commitments. In conditions (6–8), the agent acts morally while violating a prior *immoral* commitment (Rodrigo gives the money to Alenka *despite* having earlier judged it better, or having resolved, to wait for her to die and inherit her savings). Thus, while conditions (1–4) are cases of immoral action, conditions (6–8) are cases of immoral prior commitments. Condition (5) is the only case in which there is no immoral element to Rodrigo's character. The fact that intuitive participants produced higher WWAs in all conditions but condition (5) suggests that they were influenced, not only by the immorality of Rodrigo's action, but

also by the immorality of Rodrigo's prior commitments. We note that our design is not suited to test this interpretation. We did no treat the moral valence of commitments as a variable. We do emphasize, however, that our results suggest it, and that it would be adequate to consider it in the design of future studies.

Hence the evidence seems to suggest that, while intuitive participants attribute high WWA to any case that includes an immoral element (be it an immoral action or an immoral commitment), the WWAs of reflective participants are unaffected by moral value, and based solely on the violation of a specific kind of prior commitment: resolutions.

Two concepts of weakness of will

The analysis thus suggests a form of concept pluralism. We are in fact either dealing with a concept that has two distinct structures (Laurence & Margolis 1999), or with two concepts, each with a different structure (Weiskopf 2009). The intuitive and reflective groups seem to be using either two different structures of the same concept of WEAK WILL, or two structurally heterogeneous WEAK WILL concepts, given that each group performs different, and inconsistent, categorizations.

Against this reading, it is also possible to say that both groups are using one and the same WEAK WILL concept, with one and the same conceptual structure, but that one of the groups categorizes agents differently because influenced by irrelevant or concept-extraneous features. For instance, it is possible that all participants have the same concept of WEAK WILL (say, 'A person who violates prior commitments has a weak will'), but that intuitive participants are additionally influenced by a concept-extraneous element: their negative evaluation of Rodrigo's character (due either to his immoral action or to his immoral commitments). This external element's influence would lead intuitive participants to provide higher WWAs for Rodrigo whenever they evaluate him negatively. Thus, although the WWAs of reflective participants are different (because they remain unaffected by

negative evaluations of Rodrigo's actions or commitments), both groups could be using one and the same concept, with one and the same structure.[1]

Thus, there are two main possible interpretations: a dual-concept interpretation (each group applies a different concept, or a different structure of a single concept), and a common-concept interpretation (both groups apply the same concept, but one of them is also influenced by concept-extraneous elements). The evidence, however, does not support the latter view. For if there was a common concept, we should be able to infer some definitional element of the common concept from the WWAs of both groups. But what would that definitional element be?

There does not seem to be one, if we take at face value the direction in which our data are pointing. As we conjectured in the previous section, intuitive participants attribute high WWA to any case that includes an immoral element (be it an immoral action or an immoral commitment), whereas reflective participants are unaffected by moral value, and base their WWA solely on the violation of resolutions. A common definitional element appears to be absent. Rather, the intuitive group wields a concept of WEAK WILL with a core trait along the lines of *negative evaluation*:

> (WW_1) A weak will is a bad will;

whereas the reflective group employs a concept of WEAK WILL with a core trait along the lines of *resolution violation*:

> (WW_2) If a person acts in a way that violates a prior resolution, (s)he is weak-willed.

WW_1 has a prototypical structure based on an association between weakness of will on the one hand, and either acting immorally or having immoral commitments (judgment or

[1] We thank two anonymous reviewers for raising this issue.

resolution or both) on the other, so that the WEAK WILL concept applies whenever either of these two conditions is met. Accordingly, intuitive participants do not focus on violations, but on any aspect of the actions and/or commitments that elicits a negative evaluation. Just like people who go for their first intuition in the CRT, intuitive participants follow a heuristic, intuition-based categorization process based on this association between bad will and weak will.

WW_2 is a different concept[2] that takes the violation of a resolution, independently of its moral valence, as a necessary and sufficient criterion for its application. Reflective judges treat resolutions as the sole relevant commitment (perhaps because they take resolutions to reliably reflect all-things-considered judgments), and they also treat moral valence as irrelevant (perhaps because they consider moral valence and strength of will to be independent evaluative registers), so that neither violation of best judgment nor moral valence warrants categorization of the agent as weak-willed. Just like those who refrain from following their first intuition in the CRT, and check their answers by means of effortful, rule-based processes, reflective participants refrain from following the intuitive, heuristic classification of all bad wills as weak wills, and engage their reflection to assess whether a bad will may also be strong. This requires treating WEAK WILL as a concept based on a classification rule, not merely on an association.

This explains why differences in cognitive disposition are correlated with differences in weakness of will attribution. Most participants tend to base their WWA on an automatic, heuristic process based on the association between negatively-valenced terms like weakness of will and immorality. If this is the case, then participants would give higher WWA to all conditions except condition (5), because all other conditions have some element that calls for a negative evaluation (be it immoral action or immoral commitment). From

[2] We concede as possible to say that it is one and the same concept with two different core structures. In this case, despite the different structures, the concept is used invariably to make a negative evaluation of a person's character, though supported in each case in different features of her character.

this perspective, a bad will is a weak will, and a weak will is a bad will. Only reflective participants are disposed to take distance from this heuristic categorization, and subject it to critical evaluation.

5. Conclusions

No previous study of folk attributions of weakness of will included all three variables studied here, and no other study had previously measured the cognitive dispositions of participants. Through this design our study offers good evidence towards answering the question regarding the relative roles of judgment violation and resolution violation in WWA. For one, our evidence suggests that the results from May & Holton (2012), which point to the cluster character of the folk concept of weakness of will, may be an artifact of pooling together two different types of responses coming from intuitive and reflective participants. Instead, we find evidence for the claim that intuitive and reflective participants have different concepts (or engage different structures of one same concept) of weakness of will, each with a seemingly different structure. The intuitive participants treat the concept as disjunctive (*à la* Mele 2010). More specifically, when the agent acts immorally, it does not seem to matter much for intuitive participants whether the agent violated or complied with prior commitments. The scores of WWA are not significantly different in any of the conditions with immoral action. But when the agent acts morally, the immoral character of any commitment (be it judgment, resolution, or both) significantly increases WWA. In contrast, participants in the reflective group seem to use resolution violation as the sole criterion to attribute weakness of will. Holton's (1999) original intuitions seem to have captured the reflective group's understanding of the concept.

The difference in the concept of weakness of will as used by the intuitive and reflective groups is evidenced by the differential significance of moral valence in WWA. Intuitive participants attribute significantly less weakness to an agent who does not violate any prior commitments *only if* his action is moral, but make similar WWA to an agent who performs an immoral action, regardless of whether he violates a prior commitment or not. In contrast, reflective participants are impervious to the influence of the action's or the prior commitments' moral valences, and attribute significantly less weakness of will to an agent

who does not violate any prior commitments, regardless of whether the action, or the prior commitments, are moral or immoral.

All in all, our study provides strong evidence for the claim that differences in the cognitive dispositions of participants result in a differential understanding of the folk concept of weakness of will. This is also a cautionary note for future experimental studies into other folk concepts of interest to philosophers. Speaking about *the* folk concept of a philosophical term may be too hasty, since cognitive differences between participants may lead them to approach thought experiments differently, and handle the underlying philosophical concepts in diverse ways, or apply concepts with different structures. To what extent philosophical analysis should be responsive to folk concepts remains a matter of open debate, but whatever the result of that debate is, the analysis of folk concepts itself would benefit from taking cognitive dispositions into account, for it may help in avoiding unwarranted generalizations.

References

Baddeley, A. (2007). Working memory, thought, and action. Oxford University Press.

Doucet, M. & Turri, J. (2014). Non-psychological weakness of will: self-control, stereotypes, and consequences. *Synthese,* 191(16), 3935-3954.

Evans, J. S. B. T. (2010). Thinking Twice: Two minds in One Brain. Oxford University Press.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. Perspectives on Psychological Science, 8(3), 223–241.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives,* 19(4), 25–42.

Holton, R. (1999). Intention and weakness of will. *The Journal of Philosophy*, 96(5), 241-262.

May, J. & Holton, R. (2012). What in the world is weakness of will? *Philosophical Studies* 157(3), 341–360.

Mele, A. R. (1995). *Autonomous agents: From self-control to autonomy*. Oxford University Press.

Mele, A. (2010). Weakness of will and akrasia. *Philosophical Studies*, 150 (3), 391–404.

Sinayev, A. and Peters, E. (2015) Cognitive reflection vs. calculation in decision- making. Frontiers in Psychology 6:532. doi: 10.3389/fpsyg.2015.00532

Sousa, P. and Mauro, C. (2014). The evaluative nature of the folk concepts of weakness and strength of will. *Philosophical Psychology,* 28(4), 487–509.

Weiskopf, D. A. (2009). The plurality of concepts. Synthese, 169(1), 145–173.

Appendix

We append the Spanish version of the vignette and its 8 variations. The parts that change from vignette to vignette are given in brackets.

Por favor lea atentamente la siguiente historia (Please read the following story attentively):

> Rodrigo y Alenka son amigos desde la niñez y viven en la misma casa. Rodrigo estudia y Alenka trabaja. Rodrigo acaba de ganar una buena suma de dinero en un concurso, pero aún no se lo ha contado a Alenka. Desafortunadamente, ella acaba de enterarse de que sufre de una enfermedad mortal, cuyo único tratamiento efectivo es muy costoso. Sus ahorros no alcanzan. Ella le comenta esto a Rodrigo y él se da cuenta de que la suma que él ha ganado completaría el costo del tratamiento. Sabe, además, que Alenka no tiene parientes vivos, y que si ella muere, él heredará sus ahorros. Debe tomar una decisión en 24 horas y avisar al hospital, pues el tratamiento debe comenzar de inmediato y hay que pagar por anticipado.
>
> Después de sopesar las alternativas, concluye que lo mejor que puede hacer es (esperar que Alenka muera y a heredar sus ahorros/ayudar a Alenka).

(Entonces/Sin embargo) resuelve (esperar a que muera/ayudarla). (En efecto/Sin embargo) cuando llega el momento de avisar al hospital, (calla sobre el dinero del que dispone, y espera a que Alenka muera para heredar sus ahorros/anuncia que va a ayudarla y usa el dinero de su concurso para pagar el tratamiento por anticipado).

¿Qué tan de acuerdo está Ud. con la siguiente afirmación?

"En esta historia, Rodrigo demuestra una voluntad débil." **Marque un número del 0 al 10**

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
100% Desacuerdo		60% Desacuerdo		20% Desacuerdo		20% Acuerdo		60% Acuerdo		100% Acuerdo

Table 1: Mean WWA for intuitive participants by condition (vignette)

Vignette		N	Range	Mean	Std. Deviation
IM¬V	WeakW	35	10	4.94	4.284
	Valid N (listwise)	35			
IM.JV	WeakW	38	10	6.05	3.827
	Valid N (listwise)	38			
IM.RV	WeakW	39	10	5.10	3.575
	Valid N (listwise)	39			
IM.JV&RV	WeakW	26	10	5.81	3.383
	Valid N (listwise)	26			
MO¬V	WeakW	41	10	2.24	3.200
	Valid N (listwise)	41			
MO.JV	WeakW	32	10	5.41	3.600
	Valid N (listwise)	32			
MO.RV	WeakW	38	10	5.82	3.220
	Valid N (listwise)	38			
MO.JV&RV	WeakW	37	10	5.27	3.717
	Valid N (listwise)	37			

Table 2: Mean WWA for reflective participants by condition (vignette)

Vignette		N	Range	Mean	Std. Deviation
IM¬V	WeakW	26	10	3.42	3.982
	Valid N (listwise)	26			
IM.JV	WeakW	26	10	4.19	3.990
	Valid N (listwise)	26			
IM.RV	WeakW	27	10	5.70	4.017
	Valid N (listwise)	27			
IM.JV&RV	WeakW	35	10	6.26	3.576
	Valid N (listwise)	35			
MO¬V	WeakW	25	10	3.36	3.290
	Valid N (listwise)	25			
MO.JV	WeakW	31	10	3.94	3.511
	Valid N (listwise)	31			
MO.RV	WeakW	27	10	5.26	3.601
	Valid N (listwise)	27			
MO.JV&RV	WeakW	30	9	6.50	2.813
	Valid N (listwise)	30			