The Evaluative Nature of the Folk Concepts of Weakness and Strength of Will

Paulo Sousa [a, c] & Carlos Mauro [b, c] *

(forthcoming in *Philosophical Psychology*)

[a] *Institute of Cognition & Culture, Queen's University, Belfast*

[b] *School of Economics and Management, Portuguese Catholic University, Porto*

[c] *Porto X-Phi Lab, Porto*

Abstract: This article examines the evaluative nature of the folk concepts of weakness and strength of will and hypothesizes that their evaluative nature is strongly connected to the folk concepts of blame and credit. We probed how people apply the concepts of weakness and strength of will to prototypical and non-prototypical scenarios. While regarding prototypical scenarios the great majority applied these concepts according to the predictions following from traditional philosophical analyses, when presented with non-prototypical scenarios, people were divided. Some, against traditional analyses, did not apply these concepts, which we explain in terms of a clash of evaluations involving different sorts of blame and credit. Others applied them according to traditional analyses, which we explain in terms of a disposition to be reflective and clearly set apart the different sorts of blame and credit involved. Still others applied them in an inverse way, seemingly bypassing the traditional components resolution and best judgment, which we explain in terms of a reinterpretation of the scenarios driven by an assumption that everyone knows deep inside that the best thing to do is to act morally. This division notwithstanding, we claim that our results are largely supportive of traditional analyses (qua analyses of folk concepts).

Keywords: weakness of will, strength of will, blame, credit, folk psychology, evaluative judgments

·Authorship is equal.

1. Introduction

> John has been smoking for ten years. He is now in conflict because although he enjoys smoking he is preoccupied with his health. After considering all relevant aspects of the matter, John concludes that the best thing for him to do is to quit smoking. Accordingly, he decides that the next day he will quit smoking. The next day, while still thinking that it is better for him to quit smoking, John succumbs to temptation—he continues to smoke.

Weakness of will has been a theme of philosophical discussion since ancient philosophy (see Bobonich & Destrée, 2007; Charlton, 1988; Hoffmann, 2008). Most philosophical inquiry concerns whether the phenomenon of weakness of will is in fact possible and, if so, whether it violates some important standard of practical rationality. Some claim that weakness of will is impossible (e.g., Hare, 1952, 1963); others claim that it is possible albeit inherently irrational (e.g., Davidson, 1970); still others claim that sometimes it is even rational (e.g., Arpaly, 2000). We will not take a stand on these issues because this paper has a different focus. Our topic is the folk concept of weakness of will (plus its semantic complement, the folk concept of strength of will). Thus, our inquiry concerns the structure of the concept utilized by ordinary people to categorize the above scenario in terms of weakness of will (alternatively, if in the above scenario John had quit smoking not succumbing to the temptation to smoke, in terms of strength of will).[1]

In three recent articles, Alfred Mele (2010), Joshua May & Richard Holton (2011) and James Beebe (submitted) discuss the folk concept of weakness of will with the support of some empirical evidence. Their primary aim is to probe the role of two components in the structure of

[1] We presume that ordinary people in western societies categorize some episodes of practical decision-making as displaying weakness or strength of will, although we don't have exhaustive cross-cultural evidence to claim that these are universal folk concepts—there may be societies where people do not possess these concepts.

such a concept—the notions of violation of a best judgment (best judgment understood as a judgment about the course of action that is best for one to take, in light of the relevant aspects of the matter) and violation of a resolution (resolution understood as a decision that goes against some strong contrary motivational force), as we illustrated in our initial scenario.[2] For the most part, we shall leave aside the issue of the relative importance of these components and shall treat them as if they were a single one. Our primary aim instead is to investigate the evaluative nature of the folk concepts of weakness and strength of will.

Some concepts are purely evaluative—they indicate nothing more than a negative or positive value (e.g., *good* and *bad*). Some concepts are purely descriptive—they do not indicate any negative or positive value (e.g., *rock*). The folk concepts of weakness and strength of will are not purely descriptive, since they indicate a negative attribute of the agent, for weakness of will, and a positive attribute of the agent, for strength of will. This might seem trivially the case, but if so, it would entail that one would not easily apply these concepts when the agent ends up doing something considered to have a valence opposite to the valence implied by these concepts. For example, take the following scenario:

> John has been smoking for ten years. He is now in conflict because although he enjoys smoking he is preoccupied with his health. After considering all relevant aspects of the matter, John concludes that the best thing for him to do is to continue to smoke because he values immediate pleasure more than long-term health. Accordingly, he decides that

[2] The notion of best judgment at stake here is internalist in that it concerns the practical option considered to be overall better from the perspective of the agent (e.g., from John's point of view in the above scenario). In Holton's view, the decision that constitutes a resolution is often not made in the face of a strong contrary motivation, but in anticipating that a strong contrary motivation will arise in the future (see Holton 1999).

> he will continue to smoke. The next day, while still thinking that it is better for him to continue to smoke, John is swayed by his health preoccupations and quits smoking.

If the concept weakness of will implies a negative attribute and one considers that quitting smoking is good, one should hesitate in attributing weakness of will to John in this scenario because the positive evaluation implied by his action clashes with this attribution (alternatively, if in the above scenario John had continued to smoke without being swayed by his health preoccupations, one should hesitate in attributing strength of will to him for similar reasons).

In this article, we report three studies testing our prediction that ordinary people hesitate in applying the concepts of weakness and strength of will to scenarios that are susceptible to a clash of evaluations, such as the two versions described in the previous paragraph. For traditional philosophical analyses, these scenarios constitute straightforward instances of weakness and strength of will, for they fit the traditional analysans of the concepts of weakness and strength of will. For this reason, our evidence will also test the alternative predictions following from traditional philosophical analyses.

Before moving to the next section, however, let us qualify the import of our claims about the predictions of traditional philosophical analyses. First, traditional analyses of the concepts of weakness and strength of will assume that these concepts incorporate a normative aspect and, insofar as norm violation and norm compliance are deemed to have negative and positive value, respectively, an evaluative aspect too. In most analyses, an agent displays weakness of will only if she *violates her best judgment* (strength of will only if she *follows her best judgment*). In some analyses (in particular, Holton's), an agent displays weakness of will only if she *violates a resolution that she ought to stick to* (strength of will only if she *follows a resolution that she*

ought to stick to). These aspects themselves do not necessarily indicate an evaluation *of the agent*, but they imply, for traditional analyses, that the agent displays irrationality, in the case of weakness of will, and rationality, in the case of strength of will. Thus, in addition, traditional analyses suppose that the concepts of weakness and strength of will are evaluative in the sense that the application of these concepts would imply a negative attribute of the agent (the irrationality of the agent, for weakness of will) and a positive attribute of the agent (the rationality of the agent, for strength of will). For these reasons, when we described the predictions of traditional philosophical analyses as above, we are not suggesting that these analyses consider the concepts of weakness and strength of will to be purely descriptive in any relevant sense.

Second, most philosophers working on the topic have not discussed scenarios susceptible to a clash of evaluations and have not been concerned with how ordinary people apply the concepts of weakness and strength of will. Had philosophers discussed these types of scenarios with the aim of predicting how ordinary people would interpret them, they might have envisaged that people would hesitate to categorize them as instances of weakness or strength of will. For example, in relation to John quitting smoking after a resolution based on the judgment that the best thing to do is to continue to smoke, they might have predicted that ordinary people would hesitate in categorizing the scenario as an instance of weakness of will because people would consider that John displayed rationality from their externalist perspective, which clashes with an attribution of weakness of will. For this reason, when we claim that our evidence will test the alternative predictions of traditional analyses, our claim should not be understood as a claim about the predictions that philosophers have made or could have made, based on traditional analyses, on how ordinary people apply the concepts of weakness and strength of will.

Our claims are just about the predictions that logically follow from traditional analyses "in the abstract", if these analyses were taken to concern the way ordinary people apply the concepts of weakness and strength of will.

2. Basic design and predictions

Although we do not claim that the evaluative aspect of the folk concepts of weakness and strength of will is moral in nature, in our studies we focused on actions related to the moral domain (i.e., the domain of basic rights and justice) in order to make the evaluative clash more salient.[3] We presented participants with scenarios where the morality component is manipulated while the presence of a resolution based on a best judgment is kept constant. In one type of scenario, the character makes a resolution based on a best judgment that corresponds to the moral point of view (e.g., resolves not to kill based on the judgment that the best thing to do is to avoid killing). In another type of scenario, the character makes a resolution based on a best judgment that does not correspond to such a point of view (e.g., resolves to kill based on the judgment that the best thing to do is to kill). Furthermore, we manipulated whether, in acting, the character follows or does not follow the resolution based on the best judgment (e.g., whether the character kills or does not kill), in order to create situations of weakness of will, in which the character does not follow the resolution, and strength of will, in which the character does follow it. Therefore, the basic design of our studies has four types of conditions:

[3] It is important to keep in mind that in this article we use "moral" only in the specific sense of norms or evaluations concerning the domain of basic rights and justice (see Sousa, Holbrook, & Piazza, 2009), not in the general sense of normative or evaluative, which concerns any type of domain.

- The character makes a moral resolution based on a best judgment but does not follow it, thus acting immorally (e.g., resolves not to kill based on the judgment that the best thing to do is to avoid killing, but kills). We call this type of condition "MR & IA".[4]
- The character makes a moral resolution based on a best judgment and follows it, thus acting morally (e.g., resolves not to kill based on the judgment that the best thing to do is to avoid killing, and does not kill). We call this type of condition "MR & MA".
- The character makes an immoral resolution based on a best judgment but does not follow it, thus acting morally (e.g., resolves to kill based on the judgment that the best thing to do is to kill, but does not kill). We call this type of condition "IR & MA".
- The character makes an immoral resolution based on a best judgment and follows it, thus acting immorally (e.g., resolves to kill based on the judgment that the best thing to do is to kill, and kills). We call this type of condition "IR & IA".

According to traditional analyses as we discussed above, there should be no asymmetry between MR & IA and IR & MA—when the character does not follow a resolution based on a best judgment, people should be of the opinion that the character displays weakness of will. Likewise, there should be no asymmetry between MR & MA and IR & IA—when the character follows a resolution based on a best judgment, people should be of the opinion that the character displays strength of will. We agree with the traditional predictions for MR & IA and MR & MA. However, we make alternative predictions for IR & MA and IR & IA: when the character does not follow an immoral resolution based on a best judgment, thus acting morally, most people will not consider that the character displays weakness of will because the negative evaluation implied

[4] "M", "I", "R" and "A" stand for "moral", "immoral", "resolution based on a best judgment" and "action", respectively.

by the concept weakness of will clashes with the positive evaluation implied by the character's moral action; when the character follows an immoral resolution based on a best judgment, thus acting immorally, people will not consider that the character displays strength of will, because the positive evaluation implied by the concept strength of will clashes with the negative evaluation implied by the character's immoral action. Let us call the conditions MR & IA and MR & MA the "prototypical conditions," the conditions IR & MA and IR & IA the "non-prototypical conditions," and the presumed asymmetry in judgments related to the evaluative contrast between these two groups of conditions the "evaluative asymmetry." In the following three sections, we provide evidence testing the existence of such evaluative asymmetry.

3. Study one

In this study, we manipulated the aforementioned types of conditions in two different contexts—one where the character is a professional assassin, another where he is a professional robber. We hypothesized that the evaluative asymmetry would occur similarly in both contexts.

3.1. Method

3.1.1. Participants

For the professional assassin context, the participants were 82 adults living in Porto, Portugal. Of these, 69% were female. This sample did not include students and was heterogeneous in terms of education. For the professional robber context, the participants were

80 undergraduate students taking introductory courses in economics and management at the Catholic University of Portugal, Porto.[5] Of these, 34% were female. Participants from both samples were fluently Portuguese-speaking Europeans.[6]

3.1.2. Procedure, design and materials

Participants answered the questionnaire in a silent room, each seated at a separate desk. In a 2 (assassin vs. robber) x 2 (moral vs. immoral resolution) x 2 (moral vs. immoral action) between-subjects design, participants were assigned to one of the eight conditions of the study:[7]

PROTOTYPICAL CONDITIONS

John is a professional assassin (*robber*). He has started to think about quitting this profession because he feels that it is wrong to kill (*steal from*) another person. However, he is strongly inclined to continue with it because of the financial benefits.

John is in conflict, but after considering all aspects of the matter, he concludes that the best thing for him to do is to quit his profession. Accordingly, he decides that the next day he will look for a job that does not involve violence (*for an honest job*).

IA: The next day, while still completely sure that the best thing for him to do is to look for a job that does not involve violence (*for an honest job*), John is swayed by the

[5] The Catholic University of Portugal is not a religious institution.

[6] Given that we used a different sample for each context and random assignment occurred only within contexts, strictly speaking we have two independent studies here. Because there wasn't any apparent difference in terms of the results of each context and for the sake of simplicity, we treat the two studies as a single one.

[7] In this and the following studies, we phrased the scenarios in terms of the verb "decide" ("decidir" in Portuguese) instead of "resolve" ("resolver" in Portuguese), supposing that, given the explicit type of decision-making conflict portrayed in the scenarios, the decisions would be understood more specifically as resolutions. For access to the original scenarios in Portuguese, contact one of the authors.

financial benefits. Against what he had decided, he kills (*steals from*) another person for money.

MA: The next day, completely sure that the best thing to do is to look for a job that does not involve violence (*for an honest job*), John is not swayed by the financial benefits. As he had decided, he looks for a job that does not involve violence (*for an honest job*).

NON-PROTOTYPICAL CONDITIONS

John is a professional assassin (*robber*). He thinks that the financial benefits give him a good reason to continue with this profession. However, he has been strongly inclined to quit it because he feels that it is wrong to kill (*steal from*) another person.

John is in conflict, but after considering all aspects of the matter, he concludes that the best thing for him to do is to continue with his profession. Accordingly, he decides that the next day he will kill (*steal from*) another person for money.

MA: The next day, while still completely sure that the best thing for him to do is to kill (*steal from*) another person for money, John is swayed by the feeling that it is wrong to kill (*steal from*). Against what he had decided, he looks for a job that does not involve violence (*for an honest job*).

IA: The next day, completely sure that the best thing to do is to kill (*steal from*) another person for money, John is not swayed by the feeling that it is wrong to kill (*steal from*). As he had decided, he kills (*steals from*) another person for money.

In the conditions where the character does not follow his resolution, participants were asked whether they agreed or disagreed with the following statement: “John displays weakness

of will when, the next day, he kills (*steals from*) another person for money" [or, for the non-prototypical conditions, "looks for a job that does not involve violence (*for an honest job*)"]. In the conditions where the character follows his resolution, participants were asked whether they agreed or disagreed with the following statement: "John displays strength of will when, the next day, he looks for a job that does not involve violence (*for an honest job*) [or, for the non-prototypical conditions, "kills (*steals from*) another person for money]. Participants answered on a seven-point scale from strongly disagree to strongly agree, with the midpoint as neither agree nor disagree. Participants were also asked to justify their specific responses, which they did by writing down their justifications.

3.2. Results

Approximately twenty participants were in each of the eight conditions of the study. Tables 1 and 2 represent the means and distributions of the weakness and strength-of-will responses, respectively. Each table contains the prototypical and non-prototypical conditions related to its dependent measure. For computing the means, the responses on the seven-point scale were coded from -3 (strongly disagree) to 3 (strongly agree). In reporting the distributions, we reduced the three levels of disagreement and agreement to disagreement and agreement tout court not only to simplify things but also to solve a methodological problem that seems to exist with this type of scale as far as the folk concepts of weakness and strength of will are concerned. It is evident that these concepts have a graded structure, namely, that the properties denoted by them are deemed to be instantiated in different degrees—participants often qualified their justifications by saying that John displayed *huge* strength of will, *some* weakness of will etc. The

justifications given by participants also suggest that some participants incorrectly interpreted the agreement scale as a way of measuring the degree to which these properties are instantiated instead of a way of measuring the level of agreement concerning the instantiation of these properties, whatever their degree of instantiation. In this case, for example, a participant who completely agreed that John displayed a small degree of weakness of will may have marked "1" instead of "3" because of an incorrect interpretation of the agreement scale.

Table 1
Means and distributions of weakness-of-will responses

Weakness of Will		***M***	**Agreed**	**Neutral**	**Disagreed**
MR & IA	Assassin	1.2	74%	0%	26%
	Robber	2.4	95%	5%	0%
IR & MA	Assassin	-1.3	25%	10%	65%
	Robber	-1.0	30%	5%	65%

Note. MR = moral resolution/best judgment; IR = immoral resolution/best judgment; MA = moral action; IA = immoral action

Table 2
Means and distributions of strength-of-will responses

Strength of Will		***M***	**Agreed**	**Neutral**	**Disagreed**
MR & MA	Assassin	1.0	60%	20%	20%
	Robber	1.4	75%	15%	10%
IR & IA	Assassin	-1.2	20%	5%	75%
	Robber	-1.3	30%	5%	65%

Note. MR = moral resolution/best judgment; IR = immoral resolution/best judgment; MA = moral action; IA = immoral action

Our hypotheses received support from participants' responses. There is a clear evaluative asymmetry between prototypical and non-prototypical conditions in each of the dependent measures and the different contexts contributed similarly to such asymmetries. A 2 (evaluative asymmetry) x 2 (scenario context) ANOVA on each of the dependent measures showed a main effect for evaluative asymmetry in both [$F(1, 78) = 42.84$, $p < .001$, $\eta_p^2 = .36$, for weakness of will; $F(1, 76) = 24.68$, $p < .001$, $\eta_p^2 = .26$, for strength of will], but a main effect for scenario context or an interaction in none ($p > .10$ and $\eta_p^2 < .05$, for all tests).

3.3. Discussion

The low level of agreement in non-prototypical conditions does not support traditional analyses. We hypothesize that this occurred because, in IR & MA, participants see an incompatibility between the negative predicate weakness of will and the positive evaluation implied by John's following his moral judgment (i.e., by his moral action), and, in IR & IA, because they see an incompatibility between the positive predicate strength of will and the negative evaluation implied by John's not following his moral judgment (i.e., by his immoral action). However, the few agreements in IR & MA and IR & IA show that a non-negligible percentage of participants answered according to traditional analyses, which suggests that there may be individual characteristics that make one more or less sensitive to the above evaluative clashes (We return to this point in the general discussion.).

The high level of agreement in prototypical conditions is in line with what is predicted by traditional analyses—in MR & IA, participants attributed weakness of will because John did not follow his resolution based on a best judgment; in MR & MA, they attributed strength of will because he followed his resolution based on a best judgment. However, there were a few disagreements and neutral attributions in these conditions that are unexpected. Participants' written justifications thereof suggest an explanation:

In MR & IA, participants' justifications suggest that they did not attribute weakness of will in this condition because they took John to be in a "vicious situation that left him without much option in terms of a change in life-style". This rationale seems to be related to one of the components of the folk concept of weakness of will that has not been part of our main concern in

this paper—that of basic freedom. In other words, the folk concept of weakness of will does not seem to apply to cases where a constraining factor going beyond the power of a normal agent forces the agent to go against her resolution or best judgment—if the constraint is too strong, it is unreasonable to think that the agent displays weakness. Under this interpretation, in MR & IA, the majority could attribute weakness of will to John because it did not accept the existence of an insurmountable constraining factor. In other words, majority and minority in MR & IA differed only in terms of the interpretation of the situation.

In MR & MA, participants' justifications for the few disagreements or neutral attributions suggest that participants did not attribute strength of will because they focused on appraising John in terms of his series of actions instead of the next-day action that was the topic of the question—John's history of wrongdoing was incompatible with attributing strength of will, for these participants.

4. Study two

There is an alternative explanation for the asymmetry in the contexts of the first study, one based on the level of difficulty or effort involved in John's changing or not changing the profession he was accustomed to, instead of based on a clash of evaluations, as we have proposed. In this explanation, the low level of agreement in non-prototypical conditions occurred because, in IR & MA, participants see an incompatibility between the great effort involved in John's quitting his profession, which implies strength, and an attribution of weakness of will, and, in IR & IA, because they see an incompatibility between the lack of effort involved in John's not quitting his profession, which does not imply strength, and an attribution of strength

of will. Thus, it is possible that the asymmetry in participants' responses is not evaluative in nature. To probe the extent to which the asymmetry could be explained simply in terms of the difficulty or effort related to John's quitting his profession or not, in the second study we utilized a context where people's responses could not be driven by such rationale. Our hypothesis was that the asymmetry would not be significantly affected.

4.1. Method

4.1.1. Participants

The participants were 79 undergraduate students in introductory courses in economics and management at the Catholic University of Portugal, Porto. Of these, 50% were female. All participants were fluently Portuguese-speaking Europeans.

4.1.2. Procedure, design and materials

The procedure was the same as in the previous study. In a 2 (immoral vs. moral resolution) x 2 (immoral vs. moral action) between-subjects design, participants were randomly assigned to one of the following four conditions of the study:

PROTOTYPICAL CONDITIONS

John is a bank worker who has never acted dishonestly in his work. By pure chance, he discovers a way of diverting money without anyone noticing.

John is strongly inclined to continue acting honestly, because he feels that it is wrong to steal. However, he seriously envisages diverting money to his account because of the financial benefits.

John is in conflict, but after considering all aspects of the matter, he concludes that the best thing for him to do is to act honestly. Accordingly, he decides that the next day he will show the system failure to the bank, instead of stealing.

IA: The next day, while still completely sure that the best thing to do is to show the system failure to the bank, John is swayed by the financial benefits. Against what he had decided, John diverts money to his bank account.

MA: The next day, completely sure that the best thing to do is to show the system failure to the bank, John is not swayed by the financial benefits. As he had decided, he shows the system failure to the bank.

NON-PROTOTYPICAL CONDITIONS

John is a bank worker who has never acted dishonestly in his work. By pure chance, he discovers a way of diverting money without anyone noticing.

John seriously envisages diverting money to his account because of the financial benefits. However, he is strongly inclined to continue acting honestly, because he feels that it is wrong to steal.

John is in conflict, but after considering all aspects of the matter, he concludes that the best thing for him to do is to steal. Accordingly, he decides that the next day he will divert money to his bank account, instead of showing the system failure to the bank.

MA: The next day, while still completely sure that the best thing to do is to divert money to his bank account, John is swayed by the feeling that it is wrong to steal. Against what he had decided, John shows the system failure to the bank.

IA: The next day, completely sure that the best thing to do is to divert money to his bank account, John is not swayed by the feeling that it is wrong to steal. As he had decided, he diverts money to his bank account.

For the violating-resolution conditions, participants were asked whether they agree or disagree with the statement: "John displays weakness of will when, the next day, he diverts money to his bank account" (or, for the non-prototypical condition, "shows the system failure to the bank"); for the following-resolution conditions, whether they agree or disagree with the statement: "John displays strength of will when, the next day, he shows the system failure to the bank" (or, for the non-prototypical condition, "diverts money to his bank account"). Participants answered on the same scale of the previous study and were asked, in the same way, to justify their responses.

4.2. Results

Approximately twenty participants were in each of the four conditions of the study. Table 3 represents the means and distributions of the weakness and strength-of-will responses in the prototypical and non-prototypical conditions of the new context. Table 4 represents the means of weakness and strength-of-will responses in the prototypical and non-prototypical conditions of the three contexts probed so far.

Table 3
Means and distributions of weakness-of-will and strength-of-will responses

Honest Worker		*M*	Agreed	Neutral	Disagreed
Weakness Of Will	MR & IA	1.8	80%	5%	15%
	IR & MA	-1.0	35%	5%	60%
Strength Of Will	MR & MA	1.5	70%	20%	10%
	IR & IA	-0.8	37%	0%	63%

Note. MR = moral resolution/best judgment; IR = immoral resolution/best judgment; MA = moral action; IA = immoral action

Table 4
Means of weakness-of-will and strength-of-will responses in the different contexts

	Weakness of Will		Strength of Will	
	MR & IA	IR & MA	MR & MA	IR & IA
Professional Assassin	1.2	-1.3	1.0	-1.2
Professional Robber	2.4	-1.0	1.4	-1.3
Honest Worker	1.8	-1.0	1.5	-0.8

Note. MR = moral resolution/best judgment; IR = immoral resolution/best judgment; MA = moral action; IA = immoral action

Our hypothesis received support from participants' responses. There is a clear evaluative asymmetry between prototypical and non-prototypical conditions of the new context in each of the dependent measures. One-way ANOVAs on each of the dependent measures showed a significant effect for evaluative asymmetry in both [$F(1, 38) = 75.63$, $p < .001$, $\eta_p^2 = .33$, for weakness of will; $F(1, 37) = 10.89$, $p < .01$, $\eta_p^2 = .23$, for strength of will]. Furthermore, comparing the results of the two studies, the new context did not significantly change the evaluative asymmetry. The 2 (evaluative asymmetry) x 3 (scenario context) ANOVAs on the each of the dependent measures showed a main effect for evaluative asymmetry in both [$F(1, 116) = 61.21$, $p < .001$, $\eta_p^2 = .35$, for weakness of will; $F(1, 113) = 35.52$, $p < .001$, $\eta_p^2 = .24$, for strength of will], but a main effect for scenario context or an interaction in none ($p > .10$ and $\eta_p^2 < .05$ for all tests).

4.3. Discussion

The persistence of a low level of agreement in non-prototypical conditions could not be related to the level of difficulty or effort involved in John's quitting or not his profession and life-style given that this explanation does not make sense in the new context. Thus, it is plausible to suppose that this low level is mostly due to a clash of evaluations, as we have proposed. It is worth pointing out that again a non-negligible percentage of participants answered according to traditional analyses in non-prototypical conditions.

In relation to the prototypical conditions, one might have expected that disagreements or neutral attributions in MR & IA would disappear completely, for it is implausible to suppose the existence of strong constraints determining John's action in the new context. However, participants' related justifications suggest that some participants still imagined the possibility of a strong constraint dictating John's action, like the "pressure coming from his critical familial situation" or the "irresistible opportunity given to him".

One might have also expected that disagreements or neutral attributions in MR & MA would disappear, for the new context does not allow an explanation in terms of John's history of previous wrongdoing in the way we discussed before. Participants' related justifications suggest that some participants still did not attribute strength of will for a different reason—John's action was not supererogatory ("John did no more than his obligation", "John just did the right thing"). Interestingly, this rationale suggests a contrast between the folk concepts of strength and weakness of will—while strength of will has to do with actions that are supererogatory, weakness of will has to do with actions that are forbidden. It is worth noticing that, in principle, these suggestions are compatible with the fact that in MR & MA the majority considered that

John displayed strength of will, since there is important individual variation in judging the level of sacrifice that should be engaged in an action for it to be considered supererogatory instead of simply obligatory—while some people think that the actions of saints, which display huge sacrifices, are simply obligatory, other people may think that an action like John's showing the system failure is somewhat supererogatory, given the sacrifice of choosing not to profit from such 'great' opportunity.

5. Study three

In the first two studies, there was something puzzling in the participants' written justifications related to disagreements in non-prototypical conditions that suggests an even more radical departure from traditional philosophical analyses. Unexpectedly, when justifying a disagreement concerning weakness of will in IR & MA, some participants added a comment that John in fact displayed strength of will, or, when justifying a disagreement concerning strength of will in IR & IA, some added a comment that John in fact displayed weakness of will. This is puzzling because it indicates that disagreements in non-prototypical conditions were informed by an inverse attribution (e.g., in IR & MA, there would have been an attribution of strength of will instead of simply the non-attribution of weakness of will). This suggests an even more radical departure from traditional philosophical analyses because, while these analyses take the components not-following or following a resolution based on a best judgment to be part of the structure of the concepts of weakness and strength of will, inverse attributions would seem to imply that these components are irrelevant.

To probe the extent to which disagreements in non-prototypical conditions were informed by an inverse attribution and hence the import of the more radical departure, in this study we utilized all previous contexts with the following response options: (i) displays weakness of will, (ii) displays strength of will, (iii) displays neither weakness nor strength of will. With these new options, we hypothesized that the moral asymmetries between prototypical and non-prototypical conditions would reemerge in all contexts in that while in MR & IA *most* people would choose the traditional response "displays weakness of will", in IR & MA *few* people would choose this response, and while in MR & MA *most* people would choose the traditional response "displays strength of will", in IR & IA *few* people would choose this response.[8] In relation to what most people would choose in non-prototypical conditions (e.g., in IR & MA, whether the puzzling inverse response "displays strength of will" or the neutral response "displays neither weakness nor strength of will"), we did not have a specific hypothesis. We only expected that, given the aforementioned inverse comments of previous studies, at least some people would choose the inverse response. Finally, given that inverse comments were absent in the justifications related to disagreements of prototypical conditions (e.g., no justification for a disagreement in MR & IA stated that John in fact displayed strength of will), we hypothesized that no one would choose the inverse response in prototypical conditions (e.g., the option "strength of will" in MR & IA).

5.1. Method

5.1.1. Participants

[8] We use the phrase "traditional response" to refer to the response predicted by traditional philosophical analyses as discussed (e.g., "weakness of will" in MR & IA and IR & MA) and the phrase "non-traditional response" to refer to the other two responses (e.g., "neither weakness nor strength of will" and "strength of will" in MR & IA and IR & MA).

The participants were 240 undergraduate students in introductory courses in economics and management at the Catholic University of Portugal, Porto. Of these, 47% were female. All participants were fluently Portuguese-speaking Europeans.

5.1.2. Procedure, design and materials

The procedure was the same as in previous studies. In a 3 (assassin vs. robber vs. honest worker) x 2 (immoral vs. moral resolution) x 2 (immoral vs. moral action) between-subjects design, participants were randomly assigned to one of the twelve conditions of previous studies. For each condition, participants were asked which of the following options best reflected their opinion on the matter: (i) John displays weakness of will when, the next day, he … (ii) John displays strength of will when, the next day, he … (iii) John displays neither weakness nor strength of will when, the next day, he … Participants were also asked to justify their responses, which they did by writing down their justifications.

5.2. Results

Twenty participants were in each of the twelve conditions of the study. Table 5 represents the distributions of responses in the different conditions. With the percentages of non-traditional responses aggregated (e.g., in professional assassin, by adding 25% to 0% in MR & IA and adding 20% to 45% in IR & MA), we tested our hypothesis concerning the evaluative asymmetry of traditional responses between prototypical and non-prototypical conditions. Comparing the

percentages of weakness-of-will responses between the first two levels of the morality factor in each of the three contexts, the evaluative asymmetry was significant in all contexts [$\chi^2(1, N = 40) = 6.46$, $p = .01$, $\phi = .40$, for professional assassin; $\chi^2(1, N = 40) = 22.46$, $p < .001$, $\phi = .75$, for professional robber; $\chi^2(1, N = 40) = 10.00$, $p < .01$, $\phi = .50$, for honest worker]. Comparing the percentages of strength-of-will responses between the last two levels of the morality factor in each of the three contexts, the evaluative asymmetry was significant in all but one context, in which it approached significance [$\chi^2(1, N = 40) = 22.56$, $p < .001$, $\phi = .75$, for professional assassin; $\chi^2(1, N = 40) = 23.02$, $p < .001$, $\phi = .76$, for professional robber; $\chi^2(1, N = 40) = 3.64$, $p = .057$, $\phi = .30$, for honest worker].

Table 5
Distribution of responses

Morality Factor	**Professional Assassin**			**Professional Robber**			**Honest Worker**		
	-1	0	1	-1	0	1	-1	0	1
(1) MR & IA	75%	25%	0%	90%	10%	0%	75%	25%	0%
(2) IR & MA	35%	20%	45%	15%	25%	60%	25%	55%	20%
(3) MR & MA	0%	10%	90%	0%	5%	95%	0%	40%	60%
(4) IR & IA	50%	35%	15%	45%	35%	20%	40%	30%	30%

Note. MR = moral resolution/best judgment; IR = immoral resolution/best judgment; MA = moral action; IA = immoral action; -1 = displays weakness of will; 0 = displays neither weakness nor strength of will; 1 = displays strength of will

Comparing the results of the three different contexts in each of the four levels of the morality factor, there was a significant effect in IR & MA and MR & MA [$\chi^2(4, N = 60) = 9.82$, $p < .05$, $\chi^2(2, N = 60) = 9.57$, $p < .01$, respectively]. Paired comparisons amongst the different contexts in IR & MA showed that what prompted the effect was the honest-worker context [$\chi^2(2, N = 40) = 5.52$, $p = .06$, for assassin vs. honest worker; $\chi^2(2, N = 40) = 5.52$, $p < .05$, for robber vs. honest worker; $p > .30$, for the remaining comparison]. Paired comparisons amongst the

different contexts in MR & MA showed a comparable point [$\chi^2(1, N = 40) = 4.80$, $p < .05$, for assassin vs. honest worker; $\chi^2(1, N = 40) = 7.03$, $p < .01$, for robber vs. honest worker; $p > .50$, for the remaining comparison].

In regard to the distribution of non-traditional responses, a significant amount of participants chose the inverse option in all contexts of the non-prototypical conditions. This pattern indicates that many of the disagreements in the non-prototypical conditions of previous studies were informed by the puzzling inverse attribution. In the prototypical conditions, on the other hand, no participant chose the inverse response. This pattern indicates that disagreements in the prototypical conditions of previous studies were not informed by an inverse attribution, as we predicted.

Finally, it is worth pointing out that again a non-negligible percentage of participants answered according to traditional analyses in non-prototypical conditions. Interestingly, a goodness-of-fit Chi-square test against chance in each of the six non-prototypical conditions showed a significant result only in one condition ($p < .05$ for IR & MA, professional robber; $p > .10$ for all other conditions), which suggests that the distribution of responses in non-prototypical conditions is roughly trimodal, that is, that the percentage of traditional responses is similar to the percentage of each of the non-traditional ones.

6. General discussion

Overall, our results indicated three patterns of attributions in regard to non-prototypical scenarios, two of them suggesting departures from traditional analyses of the concepts of weakness and strength of will (qua folk concepts). As a less radical departure, many participants

made neither a traditional nor an inverse attribution—in IR & MA and IR & IA, these participants attributed neither weakness nor strength of will. As a more radical departure, many participants made an inverse attribution—in IR & MA, these participants attributed strength of will; in IR & IA, they attributed weakness of will. The third pattern of attribution relates to the many participants who still made a traditional attribution—in IR & MA, these participants attributed weakness of will; in IR & IA, they attributed strength of will. We propose a more detailed explanation of less and more radical patterns of non-traditional attributions in turn. We also indicate that our explanation of the less radical pattern can be extended to participants' traditional attributions.

6.1. Explaining the less radical pattern

The less radical pattern of attribution can be explained in terms of a clash of evaluations. This is the explanation we have already put forward but will now develop by arguing that the clash is better understood if the conflicting evaluations are taken to be those of blame and credit, an understanding suggested by participants' written justifications in our studies.[9]

On the one hand, we hypothesize that there is a strong connection between the folk concepts of weakness and strength of will and the folk concepts of blame and credit. Given the existence of a common normative expectation that one should not display weakness of will and should display strength of will, regardless of the content of the decisions and actions involved, the predicate weakness of will implies a certain kind of blame (i.e., blame for not following a resolution based on a best judgment) and the predicate strength of will implies a certain kind of

[9] Because "blame" and "credit" are polysemous, we would like to point out that the sense of "blame" and "credit" we have in mind here is the appraisability sense of "responsibility" (see Sousa, 2009; Zimmerman, 1988).

credit (i.e., credit for following a resolution based on a best judgment).[10] In other words, the evaluative nature of the folk concepts of weakness and strength of will is due not only to the fact that they imply a kind of weakness or strength but also to the fact that they imply a kind of blame or credit.

This hypothesized strong connection is also consistent with the following two similarities between these folk concepts. First, the existence of an insurmountable constraint that goes beyond the power of the normal agent provides a radical excuse, which undermines the reasonability of attributing blame as well as weakness of will. Discussing a similar point, Holton claims that the criterion to decide whether abandoning an intention constitutes weakness of will involves an irreducible normative dimension (Holton, 1999). In our view, this aspect of Holton's normative dimension can be explicated in terms of the possibility of excuses. Indeed, it seems quite natural to use the language of blame and exculpation to talk about weakness of will:

> In the case of my broken leg, for instance, it was clearly reasonable for me to abandon my intention [to run five miles]; that is why I *could not be charged* with weakness of will in that case. (Stroud, 2008; the emphasis is ours)

[10] Things are a bit more complicated, since there are two types of *folk* concepts of weakness/strength of will. Both types concern primarily the agent (pace philosophical traditions that take weakness and strength of will to primarily concern actions). One type, related to the way we operationalized the question in our research, corresponds to an evaluation of the agent as far as a specific decision-making process and action is concerned—the agent may display weakness of will in a specific situation, for example. The second type corresponds to an evaluation of the agent in terms of a dispositional trait—the agent may be weak-willed and consequently have a tendency to display weakness of will in specific situations, for example. Given the existence of a common normative expectation that one should not be weak-willed and should be strong-willed, the predicate weak-willed implies a certain kind of bad character trait (i.e., a vice) whereas the predicate strong-willed implies a certain kind of good character trait (i.e., a virtue), regardless of content. Thus, if we were focusing on the second type of concept or if some of our participants answered our questions with the second type of concept in mind, we would have to rephrase our discussion in terms of vice and virtue instead of blame and credit. But this wouldn't be a problem because our main points about the relation between the first type of concept and blame/credit could be made about the relation between the second type and vice/virtue. Moreover, the same relation that one finds between attributions of blame/credit and attributions of vice/virtue, one finds between attributions of weakness/strength of will and attributions of a weak/strong-willed character trait—in both cases, the recurrence of the former leads one to infer the latter and an assumption of the latter augments the probability of inferring the former in specific situations.

Second, at least concerning actions related to the moral domain, there is a general homology between weakness/strength of will and blame/credit in that blame and weakness of will have to do with actions that are forbidden, while credit and strength of will have to do with actions that are supererogatory. In this respect, it is worth pointing out that in study 3, participants' written justifications invoked the supererogatory rationale in both MR & MA and IR & MA conditions of the honest-worker context, which helps explain the relatively high percentage of neither-nor responses in these conditions (see Table 5) and the peculiar results of the honest-worker context (see section 5.2). In MR & MA this rationale counted against choosing strength of will as a traditional response; in IR & MA it counted against choosing strength of will as an inverse response.

Now, on the other hand, we hypothesize that participants give John credit for doing the right thing in IR & MA and place the blame on him for doing the wrong thing in IR & IA. Thus, many participants reading the non-prototypical scenarios may have had the following entangled conflicting intuitions: in IR & MA, John is to blame (for not following his resolution based on a best judgment) and deserves credit (for his moral action); in IR & IA, John deserves credit (for following his resolution based on a best judgment) and is to blame (for his immoral action). However, the evaluation coming from John's action is deemed preponderant in the situations, which undermines the traditional attribution, and that's why many participants did not make the traditional attribution in non-prototypical scenarios. It is important to emphasize that we are not claiming that these conflicting evaluations involve a contradiction. They are not contradictory because the evaluations are about different things—blame/credit *for not following/following a resolution based on a best judgment* vs. blame/credit *for an immoral/moral action*.

Our explanation predicts that if the blame or credit coming from the character's action were less significant to participants, it would become easier to make the traditional attribution. For example, imagine that the participants in our study were John's comrades (i.e. other professional killers or robbers). Since there would be much less significance in giving John credit for his moral action in IR & MA or in placing the blame on him for his immoral action in IR & IA, these participants would easily make the traditional attribution in non-prototypical conditions. Moreover, since setting the different evaluations apart demands a more reflective attitude, our explanation predicts that individuals with certain dispositional traits, such as being high in "need for cognition" (Cacioppo, Petty, & Kao, 1984) or having a more reflective-thinking style (Frederick, 2005), would tend to make the traditional attribution in non-prototypical conditions, even when the blame or credit coming from the character's action are significant.

This last prediction can illuminate participants' traditional attributions in non-prototypical conditions (i.e., the third pattern of attributions). It might be that these participants had the type of reflective disposition that could set clearly apart the different blame/credit evaluations involved, and lead to an answer according to the evaluation coming from whether John followed or did not follow a resolution based on a best judgment, that is, lead to a traditional attribution.

6.2. Explaining the more radical pattern

Our explanation in terms of clash of evaluations cannot explain the more radical pattern in non-prototypical conditions, namely, the inverse attributions. This is because it supposes that

the notion of a resolution based on a best judgment is still a fundamental part of the explanatory picture (otherwise one could not envisage a clash of evaluations), whereas inverse attributions suggest that this notion is irrelevant.

One may explain inverse attributions in terms of an attributional bias. Participants attributed weakness of will in IR & IA and strength of will in IR & MA simply because they thought that John was to blame and credit, respectively, for what he did and wanted to express this evaluation. In other words, participants' conceptual competence concerning weakness and strength of will was short-circuited in their attributions. It is possible that a few participants responded just in this way, as a strong emotional reaction to the scenarios—in particular, to IR & IA. However, we don't think that this is the best explanation for the great majority of inverse attributions. The concepts weakness and strength of will imply a certain kind of blame and credit, as we argued above, but they have no clear connection with blame and credit for what someone does. Thus, it is implausible that participants were using these concepts to express blame and credit for what someone does.

This first type of explanation of inverse attributions claims that the component resolution based on a best judgment is bypassed in non-prototypical scenarios. A second type of explanation claims that this component is considered to be irrelevant in both non-prototypical and prototypical scenarios (an explanation suggested to us by Joshua Knobe, personal correspondence). Judgments of weakness and strength of will might concern whether a person's action is aligned with her essential will, namely, the desires coming from her true self (not whether the action is an aligned with a reflective mental state such as a resolution based on a best judgment). Moreover, the content of this essential will may be always interpreted in 'moral' terms, and according to the 'moral' values of who is attributing weakness or strength of will.

Therefore, because it is plausible to suppose that the participants in our studies share the view that it is morally wrong to kill or steal, they understood John's essential will as the desire to avoid killing or stealing. Accordingly, participants might have attributed weakness of will in MR & IA and IR & IA because they thought that John's actions did not align with his essential will, and they might have attributed strength of will in MR & MA and IR & MA because they thought that his actions aligned with it.

We see many problems with this explanation. First, although it is undeniable that one of the ordinary meanings of the word "will" is *desire*, the ordinary meaning of "will" in the expressions "weakness of will" and "strength of will' does not seem to be reducible to desire tout court, even less so to essential desire (the same can be said in relation to the meaning of "vontade" in Portuguese, the word that corresponds to "will" in English). What is considered to be weak or strong when one uses the expressions "weakness and strength of will" is something more like the reflective decision related to a resolution based on a best judgment. So, this proposal seems to depart from the ordinary meaning of the expressions "weakness of will and strength of will" and, consequently, from the folk concepts of weakness and strength of will. Second, participants' justifications of their traditional responses related to the prototypical *and* non-prototypical scenarios of all our studies suggest that the notion of a resolution based on a best judgment is relevant—for example, participants said things like "John did not follow his decision", "John succumbed to temptation",[11] "John did not follow what he thinks was best to do" to justify an attribution of weakness of will, and sometimes in the same justification. Third, other results coming from the current literature indicate that the component resolution based on a best judgment is deemed important and it is doubtful that Knobe's proposal can explain the range

[11] The expression "succumb to temptation" ("cair na tentação", in Portuguese) was used even in the justifications related to non-prototypical scenarios, in which case the temptation was supposed to be the moral preoccupation.

of current results (see Mele, 2010; May & Holton, 2011; Beebe, submitted). Finally, this explanation in itself is difficult to conciliate with the fact that many participants made traditional attributions in non-prototypical conditions and many participants made neither traditional nor inverse attributions in these conditions.

Even if this specific explanation proposed by Knobe is unconvincing, one may still suppose that the more radical departure from traditional analyses evinced by inverse responses constitutes a clear instance of the Knobe effect in that it illustrates another case in which 'moral' considerations play a fundamental role in people's conceptual competence (see Knobe, 2010). We are skeptical about this broader perspective too. First, the idea that 'moral' considerations influenced inverse responses by playing a fundamental role in people's conceptual competence concerning weakness and strength of will could not be worthwhile pursuing without a specific characterization of the fundamental role played by 'moral' considerations—to name an effect is not to explain it.[12] Second, it is unclear to us how one could provide such specific characterization in a way that would be convincing. Third, we believe that in this broader perspective the less radical departure from traditional analyses evinced by neither-nor responses should constitute an instance of the Knobe effect too; therefore, complicating matters, one would have to provide a convincing characterization of how 'moral' considerations influenced both less and more radical departures from traditional analyses. Finally, although Knobe has discovered evaluative asymmetries in the application of a variety of folk psychological and causal concepts, we don't think that there is a general explanation for these asymmetries, let alone a general

[12] It is important to note that Knobe's specific explanation discussed above, which may be seen as an attempt to flesh out a specific characterization in this respect, postulates that the conceptual competence concerning weakness and strength of will is different from the one supposed by traditional analyses, since the component resolution based on a best judgment is not part of the competence. Thus, in order to provide a specific characterization, one needs to specify both the conceptual competence concerning weakness and strength of will and how 'moral' considerations play a fundamental role in this competence.

explanation in terms of the fundamental role of 'morality'. Take the evaluative asymmetry in attributions of intentional action, which is by far the most impressive given its effect size. The main factor in the explanation of this asymmetry has to do with the fact that the expression "intentional action" is polysemous, which is completely unrelated to the idea that 'morality' plays a fundamental role in people's conceptual competence (see Cova, Dupoux, & Jacob, 2012; Mele, 2012; Nichols & Ulatowski, 2007; Sousa & Holbrook, 2010).

For the above reasons, we would like to put forward an alternative explanation of inverse responses, one in which the component resolution based on a best judgment was important for participants' judgments in both prototypical and non-prototypical scenarios, one not akin to a Knobe-effect type of explanation. In our view, most participants with inverse attributions reinterpreted the non-prototypical scenarios; in particular, they reinterpreted the content of John's best judgment. It is not that participants left aside John's best judgment and focused simply on his moral judgment; it is rather that they did not accept the possibility of a best judgment that would go against the moral point of view—they assumed that, as everyone knows deep inside, John knew that the best thing to do is to avoid killing or stealing, and they reinterpreted John's moral judgment as his own best judgment. If so, participants attributed strength of will in IR & MA because John followed his best judgment and they attributed weakness of will in IR & IA because he did not follow it. We are aware that our point here does not concern directly the notion of resolution, but it is worth noticing that our point *does* have implications for the detection of a resolution in that many participants may have not accepted that John had made a real resolution to kill or steal, given the perception that his deep voice of conscience never let him be completely settled in killing or stealing. If so, in IR & MA, for example, participants may have attributed strength of will because John in the end changed his

mind, making a new, now real, resolution based on the his own best judgment that he should not kill or steal, and following it.

There is a kind of anomaly in the description of non-prototypical scenarios that led many participants to reinterpret them. If one's judgment about what is best to do is roughly equivalent to one's judgment about what one ought to do, and if participants reckon that deep inside everyone knows that what one ought to do in these scenarios is to avoid killing or stealing, the description of the scenarios becomes indeed anomalous. Then, participants might have reinterpreted the scenarios to make them coherent by attributing to John the judgment that the best thing to do is to avoid killing or stealing. Another point in favor of our explanation is that many of participants' written justifications in non-prototypical scenarios included comments linking John's moral judgment to his "deep convictions and values", to his "true conscience" and to "what everyone knows", which suggests that participants indeed assumed that deep inside John knew that the best thing to do was to avoid killing or stealing.

Our explanation predicts that inverse attributions in non-prototypical scenarios would decrease in contexts where people do not tend to assume universal knowledge about what is the best thing to do—e.g., in evaluative but non-moral contexts such as the one we illustrated in the introduction of this article, presumably. More to the point, it predicts a strong positive correlation between inverse attributions and such universalist assumption.

7. Conclusion

We would like to conclude with some remarks on the relationship between our results and philosophical analyses of the concepts of weakness and strength of will (qua folk concepts).

First, our results are largely supportive of traditional philosophical analyses. With regard to prototypical scenarios, the great majority of participants applied the folk concepts of weakness and strength of will according to these analyses. Moreover, of those who evinced non-traditional responses in prototypical scenarios, only the ones with the non-supererogatory type of justification for not attributing strength of will show something novel, which we interpreted in terms of an asymmetry between weakness and strength of will parallel to an asymmetry between blame and credit. On the other hand, in regard to non-prototypical scenarios, although many participants evinced non-traditional responses, one should not make too much of this evidence. Roughly one third of participants answered according to traditional analyses. Another third, we hypothesize, reinterpreted the scenarios driven by the assumption that everyone knows deep inside that the best thing to do is to act morally, and, with this reinterpretation in mind, reasoned in a way compatible with traditional analyses to arrive at inverse responses. The remaining participants indeed reasoned differently with their responses, but, if our clash-of-evaluations explanation is correct, rather than indicating something intrinsic about conceptual structure, this might show simply that, in the categorization of non-prototypical scenarios, the evaluative clash biases participants towards non-traditional responses.

Second, regarding the debate on the relative importance of the components resolution and best judgment, participants' justifications seemed to appeal to both components across all conditions, as we illustrated in our criticism of Knobe's explanation of our results. Therefore, even if it wasn't our intention to probe this debate and even if it is difficult to code and quantify participants' justifications in this respect, our results suggest that both components are part of the structure of the folk concepts of weakness and strength of will, which goes in the direction of a consensus reached in the recent debate between Mele and Holton (see introduction).

Finally, let us return to our discussion at the end of the introduction concerning the fact that philosophers could have envisaged a clash of evaluations of rationality that would lead ordinary people to avoid applying the concepts of weakness and strength of will to non-prototypical scenarios—e.g., to avoid categorizing the scenario where John quit smoking after a resolution based on the judgment that the best thing to do is to continue to smoke as an instance of weakness of will because John displayed rationality from an externalist perspective, which clashes with an attribution of weakness of will. In this respect, it is important to note that more recently, discussing scenarios akin to the non-prototypical ones discussed in this article, instead of envisaging such a clash of evaluations of rationality and the possibility that the concept of weakness of will does not apply to these scenarios, philosophers have argued that some instances of weakness of will are tout court rational (see Arpaly, 2000; McIntyre, 1990). Our results suggest that much of the literature attributing rationality to weakness of will is driven by intuitions related to the evaluative clash we have characterized. It is just that these philosophers reflectively hesitate to avoid applying the concept of weakness of will; instead, they set aside the internalist irrationality and blame expressed by someone who does not follow a resolution based on a best judgment.

Acknowledgements

The research in this article was inspired by initial research on the topic carried out by Carlos Mauro in the context of his PhD dissertation (see Mauro, 2009). We would like to thank James R. Beebe, Joshua Knobe, Nora Parren and two anonymous referees for their comments and suggestions. We would like to thank Joshua Knobe for his generous support of our research.

References

Arpaly, N. (2000). On acting rationally against one's better judgment. *Ethics*, 110, 488-513.

Beebe, J. (*submitted*). The Folk Conception of Weakness of Will.

Bobonich, C., & Destrée, P. (eds.) (2007). *Akrasia in Greek Philosophy: From Socrates to Plotinus*. Leiden, Boston: Brill.

Bratman, M. (1979). Practical Reasoning and Weakness of the Will. *Noûs*, 13, 153-171.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306–307.

Charlton, W. (1988). *Weakness of Will*. Oxford: Basil Blackwell.

Cova, F., Dupoux, E., & Jacob, P. (2012). On doing things intentionally. *Mind & Language*, *27*, 378-409.

Davidson, D. (1970). How Is Weakness of the Will Possible? In J. Feinberg (Ed.), *Moral Concepts*. Oxford: Oxford University Press.

Federick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic*

Perspectives, 19, 25-42.

Hare, R. M. (1952). *The Language of Morals*. Oxford: Clarendon Press.

Hare, R. M. (1963). *Freedom and Reason*. Oxford: Clarendon Press.

Hoffmann, T. (ed.) (2008). *Weakness of Will from Plato to the Present*. Washington: Catholic University of America Press.

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

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*, 315-365.

McIntyre, A. (1990). Is Akratic Action Always Irrational? In *Identity, Character, and Morality*, O. Flanagan and A. Rorty (eds.), Cambridge, MA: MIT Press, pp. 379-400.

Mauro, C. (2009). Uma Concepção Deflacionista da Racionalidade na Acção (A Deflationary Conception of Rationality in Action). Porto: Departamento de Filosofia - Universidade do Porto.

May, J., & Holton, R. (2010). What in the world is weakness of will? *Philosophical Studies*, 157, 341-360.

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

Mele, A. (2012). Folk conceptions of intentional action. *Philosophical Studies*, *22*, Action Theory, 281-297.

Nichols, S. and Ulatowski, J. 2007. Intuitions and individual differences: The Knobe effect revisited. *Mind and Language* 22(4): 346-365.

Sousa, P. (2009). A cognitive approach to moral responsibility—the case of a failed attempt to kill. *Journal of Cognition and Culture*, 9, 171-194.

Sousa, P., & Holbrook, C. (2010). Folk concepts of intentional action in the contexts of amoral and immoral luck. *Review of Philosophy and Psychology*, *1*, 351-370.

Sousa, P., Holbrook, C., & Piazza, J. (2009). The morality of harm. *Cognition, 113,* 80-92.

Stroud, S. (2008). Weakness of Will. *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/weakness-will/

Zimmerman, M. (1988). *An Essay on Moral Responsibility*. New Jersey: Rowman & Littlefield.