Running Header: Toxic intentions

Toxic Intentions[*]

Wesley Buckwalter

Department of Philosophy

Institute for Philosophy and Public Policy

George Mason University

wesleybuckwalter@gmail.com

John Turri

Philosophy Department

Cognitive Science Program

University of Waterloo

john.turri@gmail.com

[*] This is the penultimate version of a paper forthcoming in *Inquiry*. Please cite the final, published version if possible. Data and materials available here: https://osf.io/b8t9s/

Abstract

Pure voluntarism is the claim that we have the same voluntary control over intentions as we do decisions. The Toxin Puzzle is often taken to challenge pure voluntarism by supporting a reasons constraint on intentions. According to this constraint, one cannot voluntarily intend to do something that one lacks a practical reason to do. We present the results of three experiments stemming from this puzzle demonstrating that the concept does not support a reasons constraint and suggests that intentions are regarded as voluntary to the same degree that decisions are.

Keywords: intention; voluntarism; reasons; decision; belief

Toxic Intentions

Are intentions more like beliefs or decisions? That is, are intentions freely chosen in the same way decisions are? Or is it the case that "one cannot intend whatever one wants to intend any more than one can believe whatever one wants to believe" (Kavka, 1983, p. 36)? And if so, is it because "our intentions are constrained by our reasons for action" in much the same way that "our beliefs are constrained by our evidence" (Kavka, 1983, p. 36)?

There are two issues here. The first is whether it is conceptually possible for intention to be under voluntary control, in the same way that decisions are. Call this *pure voluntarism*. The second is whether reasons for action constrain the formation of intention, in the sense that one cannot voluntarily intend to do something that one lacks practical reason to do. Call this the *reason constraint*.

A famous and longstanding intuition suggests a pair of verdicts about these principles, namely that pure voluntarism about intention is false and the reason constraint is true. This intuition is said to arise in cases like the one first proposed by Kavka (1983), where reasons to do something separate from reasons for intending to do it:

> You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects…The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed…All you have to do is…intend at midnight tonight to drink the stuff

> tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin. (The presence or absence of the intention is to be determined by the latest 'mindreading' brain scanner...) (Kavka, 1983, pp. 33-34)

Can you intend to drink the toxin? According to Kavka, the answer is “no.” If “intentions were simply decisions, and decisions were volitions fully under the agent’s control, there would be no problem” (Kavka, 1983, p. 35). But alas they are not. Thus pure voluntarism about intention is false. Furthermore, it is false because of the reason constraint. One cannot intend to drink the toxin because one has no reason to drink it.

In this paper, we investigate whether it is conceptually possible to form the intention in the toxin case, and whether reasons for action constrain the attribution of intention. Some have suggested that it is possible to form the intention in the toxin case, if the agent is described as “bizarre” (Mele 1992). We will not avail ourselves of this strategy. Instead, we will investigate whether the ordinary intention concept, familiar to all of us from everyday folk psychology, allows for this even without specifying bizarre details.

Prior research in experimental cognitive science suggests that many mental states are regarded as voluntary, to various extents, including intention, desire, and belief (Turri, Rose and Buckwalter, 2018; Buckwalter, Rose, and Turri *in press*). Specifically, these studies suggest that, according to the ordinary concept, one can choose to believe things that are against one’s evidence in various cases. If folk psychology embraces voluntarism for belief, it is a reasonable conjecture that it might also embrace voluntarism for intention. Despite being suggestive, prior findings do not address the present research questions concerning intention and the reason constraint, because they did not compare intention to decision or test whether intentions are attributed without reasons.

To address our research questions, we conducted three behavioral experiments. The findings from our first experiment support the pure voluntarism for intention. Intention is viewed as voluntary in much the same way that decisions are. The findings from our second experiment support the conclusion that there is no reason constraint on intention. Intention is strongly attributed to agents even when reasons for action are strongly denied. The results of a third experiment demonstrate that even relatively weak-willed individuals can do this if they choose. Taken together, our findings provide strong initial evidence that folk psychology accepts pure voluntarism and rejects the reason constraint.

1 Experiment 1: the classic toxin case

1.1 Method, materials, and procedure

No research on the topic existed to inform an a priori power analysis regarding sample size, so we decided in advance to recruit approximately 50 participants per condition, plus a few extra as a precaution against attrition (see pre-registration).

We report all manipulations and measures used throughout this and the following experiments. All participants were adult residents of the United States. No participants were excluded from analysis. We recruited and tested people using an online platform of Amazon Mechanical Turk (https://www.mturk.com), TurkPrime (Litman, Robinson, and Abberbock 2017), and Qualtrics (https://www.qualtrics.com). Participants completed a brief demographic questionnaire after testing. We used R 3.5.1 for all analyses (R Core Team 2018). All stimuli, data, and code are available through the Open Science Foundation (https://osf.io/b8t9s/). All studies were pre-registered.

One hundred eleven people participated in the study. Their mean age was 34.87 years (range = 20-70, SD = 10.87), 41% (45 of 111) were female, and 94% reported native competence in English.

Participants first read a brief scenario, then responded to seven test statements. The scenario was about an agent, Greg, who was offered an opportunity of the sort envisioned in Kavka's original toxin case, with minor changes to improve readability and minimize confusion. Participants were randomly assigned to one of two conditions (refuse, choose). In the choose condition, he can earn one million dollars now, if he intends to drink a vial of toxin tomorrow. In the refuse condition, Greg refuses the challenge. In the choose condition, he accepts the challenge (see Appendix).

Beneath the scenario, on the same page, participants responded to four test statements (order rotated randomly):

Greg intends to drink the toxin. (intention)

Greg decided to drink the toxin. (decision)

Greg will enjoy drinking the toxin. (enjoyment)

Greg is excited to drink the toxin. (excitement)

Participants then proceeded to a new screen and responded to three more statements (order rotated randomly):

It was up to Greg whether he intended to drink the toxin. (voluntary)

Greg has good reason to drink the toxin. (reason)

The podcasters owe Greg the money. (owe)

Responses to all seven test statements were collected on a standard 7-point Liker scale, 1 ("strongly disagree") - 7 ("strongly agree"), arranged left-to-right on the participant's screen.

1.2 Results

We predicted that assignment to condition would affect response to the intention attribution, with attribution high in the choose condition and low in the refuse condition. Two other important questions are whether people would respond similarly to the intention and decision attributions, and whether participants would judge that the agent has good reason to drink the toxin. To evaluate these issues, we conducted a linear mixed effects analysis and followed up with appropriate t-tests. Our predictions and analysis plan were pre-registered.

The linear analysis included as fixed effects assignment to condition (between-subjects: refuse, choose), type of judgment (within-subjects: intention, decision), an interaction between condition and attribution type, and participant age and sex. It also included participant as a random factor. The only significant effect was assignment to condition (see Fig. 1 and Table 1).

Table 1
Experiment 1. Analysis of variance for the mixed linear model's fixed effects.

	Sum Sq	NumDF	DenDF	F	p
Condition	69.784	1	107	163.611	<.001
Judgment	0.004	1	109	0.01	.919
Sex	0.003	1	107	0.006	.938
Age	1.213	1	107	2.844	.095
Condition: Judgment	0.004	1	109	0.01	.919

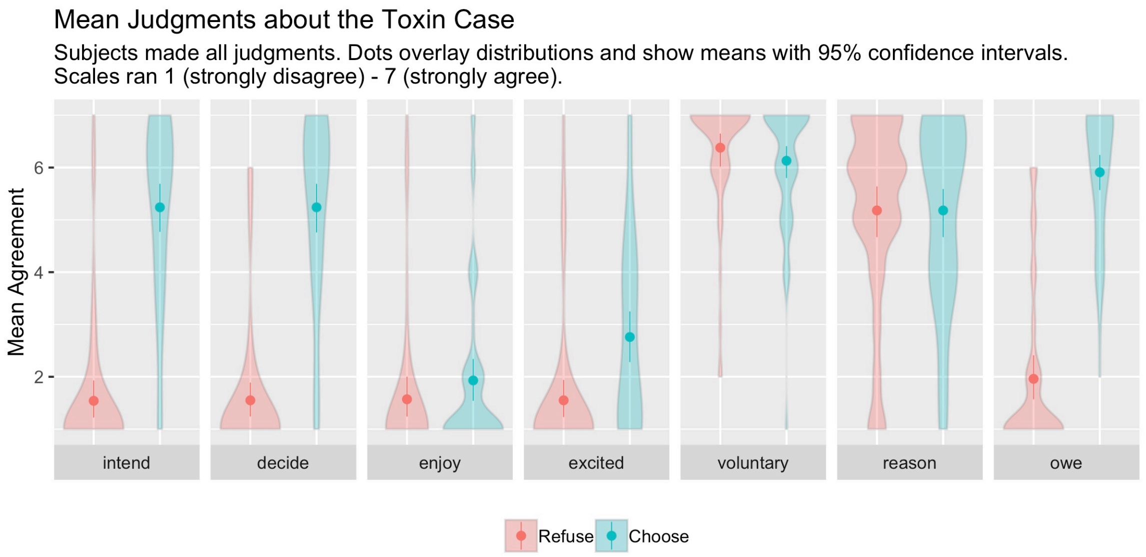

Fig 1: Experiment 1. Mean response overlaying distributions for the test statements (within-subjects) across two conditions (refuse, choose) (between-subjects). Scales ran 1 (“strongly disagree”) – 7 (“strongly agree”). Error bars show 95% bootstrapped confidence intervals.

Follow-up independent samples t-tests revealed a very large effect of condition on attributions of intention, MD = -3.7, [-4.29, -3.11], t(99.77) = -12.45, d = 2.36, and decision, MD = -3.68, [-4.29, -3.08], t(99.33) = -12.08, d = 2.29. Follow-up one sample t-tests revealed that mean intention attribution was above the midpoint in the choose condition (M = 5.24, SD = 1.77), t(54) = 5.17, p < .001, d = 0.7, and it was below the midpoint in the refuse condition (M = 1.54, SD = 1.32), t(55) = -13.96, p < .001, d = -1.87. Similarly, mean decision attribution was above the midpoint in the choose condition (M = 5.24, SD = 1.83), t(54) = 5.02, p < .001, d = 0.68, and it was below the midpoint in the refuse condition (M = 1.55, SD = 1.35), t(55) = -13.59, p < .001, d = -1.82. Reflecting that same pattern, participants judged that the agent was owed the money in the choose condition (M = 5.91, SD = 1.27), t(54) = 11.18, p < .001, d = 1.51, but that he was not owed the money in the refuse condition (M = 1.96, SD = 1.6), t(55) = -9.55, p < .001, d = -1.28. Importantly: participants judged the agent's response to be voluntary in both conditions (choose, M = 6.13, SD = 1.19, t(54) = 13.29, p < .001, d = 1.79; refuse, M = 6.38, SD = 1.18, t(55) = 15.01, p < .001, d = 2.01), and they judged that the agent had good reason to perform the action in both conditions (choose, M = 5.18, SD = 1.81, t(54) = 4.85, p < .001, d = 0.65; refuse, M = 5.18, SD = 1.81, t(55) = 4.87, p < .001, d = 0.65). Participants consistently judged that the agent would not enjoy or be excited about performing the action (see Fig. 1). Lastly, we conducted one exploratory correlational analysis and found that attributions of intention and decision were very strongly positively correlated, r(111) = 0.93, p < .001.

1.3 Discussion

This experiment examined how people judge a version of the classic toxin case. In the toxin case, an agent is offered a powerful inducement to form an intention to perform an unpleasant action. More specifically, the agent is offered one million dollars right now in exchange for intending to drink a mild toxin tomorrow. We tested two versions of the case, one in which the agent accepts the challenge by saying, "Okay, I'll do it," and one in which the agent refuses the challenge by saying, "No, I won't do it." In keeping with our prediction, people judged that the agent can intend to drink the toxin. This is proven by the fact that people judged that the agent actually forms the intention when he accepts the challenge, although they deny that he forms the intention when he rejects the challenge.

Reflecting that same basic pattern, people judged that the agent was owed the money when he accepted the challenge, but not when he rejected the challenge. Also contrary to philosophical claims about the case, people did not distinguish between forming an intention and making a decision. Overall, these results are well explained by pure voluntarism, or the hypothesis that, conceptually, intention is fully under voluntary control. Finally, and again contrary to philosophical claims about the case, people judged that the agent has good reason to drink the toxin.

This last finding raises the possibility that the toxin case is not a good candidate for illustrating the conceptual point about intention that some philosophers were interested in making. In particular, they have claimed that the agent cannot intend to drink the toxin because he doesn't have good reason to drink it. Thus it is possible that if we could convince people that the agent doesn't have good reason to drink the toxin, then they would also deny that he intends to drink it, even if he accepts the challenge. By contrast, if folk psychology embraces pure voluntarism and thus views intention as voluntary, then people will continue to agree that he intends to drink the toxin if he accepts the challenge. The next experiment directly investigates this question and thereby the status of a reason constraint on intention in folk psychology.

2 Experiment 2: choice and reasons

2.1 Method, materials, and procedure

As with Experiment 1, we again decided in advance to recruit approximately 50 participants per condition, plus some extra as a precaution against attrition. Two hundred eleven people participated in the study. Their mean age was 34.37 years (range = 19-67, SD = 10.09), 38% (81 of 211) were female, and 96% reported native competence in English.

Participants were randomly assigned to one of four conditions in a 2 (option: refuse, choose) × 2 (reason: low, high) design. Participants first read a brief scenario, then responded to seven test statements. The scenario was based on the one used in Experiment 1. The option manipulation was the same as in Experiment 1: the agent a refuses or chooses to accept the challenge. The reason manipulation was coarse and intended to affect whether the agent has good reason to accept the challenge. In particular, it was intended to make it seem that he doesn't have good reason in the low condition, and that he does have good reason in the high condition (see Appendix).

Participants then responded in the same way to the same seven test statements as in Experiment 1.

2.2 Results

We predicted that the reason manipulation would be effective, that there would be an effect of option on intention attribution, that people would continue to attribute intention in the choose low condition, that people would continue to view the intention-formation as voluntary, and that people would continue to view intention and decision similarly. In light the findings from Experiment 1, this experiment aimed to answer another critical question. If a standard interpretation of the toxin case is correct and our reason manipulation in the present experiment is effective, then we will observe a specific interaction on intention attributions: people will attribute intention in the choose high condition, but they will deny it in the other three conditions, including the choose low condition. By contrast, and in line with our predictions, if pure voluntarism is true, then people will continue to attribute intention in the choose low condition. Pure voluntarism doesn't necessarily predict no interaction or no effect of the reason manipulation. Our predictions and analysis plan were pre-registered.

We conducted a linear mixed effects analysis of participant response to the intention and decision attributions. The model included as fixed effects assignment to option (refuse, choose) and reason (low, high) conditions (between-subjects), an interaction term for option and reason, type of judgment (within-subjects: intention, decision), and participant age and sex. It also included participant as a random factor. The only significant effect was assignment to option condition (see Fig. 2 and Table 2).

Table 2
Experiment 2. Analysis of variance for the mixed linear model's fixed effects.

	Sum Sq	NumDF	DenDF	F	p
Option	293.429	1	205	390.998	< .001
Reason	1.685	1	205	2.245	.136
Judgment	2.903	1	210	3.868	.051
Sex	0.027	1	205	0.036	.851
Age	2.427	1	205	3.234	.074
Option: Reason	0.301	1	205	0.401	.527

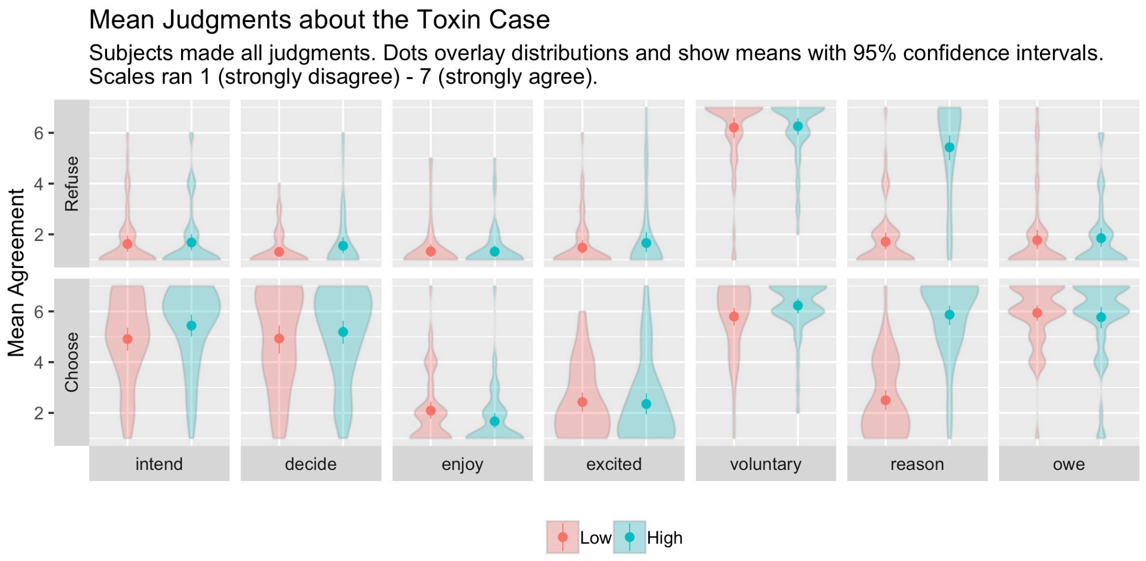

Fig.2: Experiment 2. Mean response overlaying distributions for the test statements (within-subjects) across four conditions (option: refuse, choose) (reason: low, high) (between-subjects). Scales ran 1 ("strongly disagree") – 7 ("strongly agree"). Error bars show 95% bootstrapped confidence intervals.

A follow-up independent samples t-test revealed that the effect of option was very large, t(192.4) = -17.8, p < .001, d = -2.45. Mean intention attribution was significantly above the neutral midpoint (=4) in the choose conditions, M = 5.17, SD = 1.64, t(105) = 7.34, p < .001, d = 0.71, and it was significantly below the neutral midpoint in the refuse conditions, M = 1.65, SD = 1.2, t(104) = -20.07, p < .001, d = -1.96. In the critical choose low condition specifically, mean intention attribution was significantly above the neutral midpoint, M = 4.91, SD = 1.69, t(53) = 3.96, p < .001, d = 0.54. We conducted a correlational analysis and found that attributions of intention and decision were very strongly positively correlated, r(211) = 0.86, p < .001.

We conducted an analysis of variance on participant evaluation of the agent's reasons, including as independent variables assignment to reason and option conditions, and participant age and sex (see Fig. 2 and Table 3). Assignment to reason condition, option condition, and participant age had significant effects.

Table 3
Experiment 2. Analysis of variance for reason evaluations.

	Df	Sum Sq	F	p
Option	1	16.57	7.222	.008
Reason	1	662.045	288.55	<.001
Sex	1	0.001	0	.986
Age	1	12.697	5.534	.020
Option: Reason	1	1.883	0.821	.366
Residuals	205	470.35		

A follow-up independent samples t-test for the effect of reason (low, high) revealed a very large effect, t(201.93) = -16.51, p < .001, d = -2.27. Mean reason evaluation was significantly above the neutral midpoint (=4) in the high conditions, M = 5.65, SD = 1.69, t(104 = 10.01, p < .001, d = 0.98, and it was significantly below the midpoint in the low conditions, M = 2.11, SD = 1.41, t(105) = -13.78, p < .001, d = -1.34. An independent samples t-test for the effect of option was insignificant and recorded a small effect size, t(207.16) = -1.74, p = .084, d = -0.24. Older participants tended to record lower reason evaluations. In a simple regression predicting reason evaluation from participant age, the standardized coefficient was -0.029, p = .073, suggesting that, other things being equal, an age difference of 35 years would lower reason evaluation by one point on the seven-point scale.

We conducted an analysis of variance of voluntariness judgments, including as independent variables assignment to option and reason conditions, and participant age and sex (see Fig. 2 and Table 4). The only significant effect was participant age. Older participants tended to record higher voluntariness judgments. In a simple regression predicting voluntariness judgments from participant age, the standardized coefficient was 0.025, p = .004, suggesting that, other things being equal, an age difference of 40 years would raise voluntariness scores by one point on the seven-point scale. Overall, mean voluntariness judgments were very high, M = 6.12, SD = 1.28, t(210) = , p < .001, d = 1.66.

Table 4
Experiment 2. Analysis of variance for voluntariness judgments

	Df	Sum Sq	F	p
Option	1	2.758	1.758	.186
Reason	1	3.151	2.009	.158
Sex	1	0.195	0.124	.725
Age	1	12.963	8.265	.004
Option: Reason	1	2.187	1.394	.239
Residuals	205	321.543		

2.3 Discussion

This experiment began investigating the reason constraint on intention in folk psychology and, in the process, replicated and extended the findings from Experiment 1 by testing variations of the classic toxin case. We manipulated whether the agent chose to accept the challenge (option factor), and whether the agent had good reason to accept the challenge (reason factor). An outstanding question from Experiment 1 was whether participants would continue to attribute the intention to drink the toxin, if they judged that the agent did not have good reason to drink it. The present results answer this question in the affirmative. In the choose low condition, the agent chooses to accept the challenge, even though he does not have good reason to accept it. Participants denied that he has good reason to drink the toxin, but they nevertheless judged that he intended to drink the toxin, and that he did so voluntarily. Attributions of intention were unaffected by a powerfully effective manipulation of the agent's reasons, suggesting that having a good reason for action is not a conceptual requirement for intending to perform it. As in Experiment 1, judgments about intending and deciding were very similar. Overall, these findings undermine a reason constraint on intention in folk psychology. But the pattern is well explained by pure voluntarism.

3 Experiment 3: willpower

The findings from the first two experiments raise an intriguing possibility. Existing findings suggest that belief attributions could be affected by perceived willpower (Turri, Rose and Buckwalter, 2018). Likewise, perhaps intentions are influenced by perceived willpower. The present experiment addresses this.

3.1 Method, materials, and procedure

We decided in advance to recruit approximately 100 participants per condition, plus a few extra as a precaution against attrition. Compared to the first two experiments, we decided to recruit more participants per condition because we suspected that if perceived willpower did affect intention attributions, then the effect would be small. Our suspicion was based on previous research demonstrating that the mental state of belief is conceptually voluntary (Turri, Rose, and Buckwalter 2018: Experiment 2). Even though belief is conceptually voluntary, manipulating perceived willpower had only a small, trending effect on belief attributions. Two hundred ten people participated in the study. Their mean age was 33.67 years (range = 18-70, SD = 10.33), 42% (88 of 210) were female, and 94% reported native competence in English.

Participants were randomly assigned to one of two conditions (low, high). Participants first read a brief scenario, then responded to four test statements. The scenario was based on the one used in the choose low condition of experiment 2. Despite not having good reason to accept the challenge, the agent nevertheless chooses to accept it. The conditions differed in whether the agent was described as extremely weak or strong willed (see Appendix).

Beneath the scenario, on the same page, participants responded to two test statements (order rotated randomly):

Greg intends to drink the toxin. (intend)

Greg will enjoy drinking the toxin. (enjoy)

Participants then proceeded to a new screen and responded to two more statements (order rotated randomly):

It was up to Greg whether he intended to drink the toxin. (voluntary)

The podcasters owe Greg the money. (owe)

Responses were collected on the same 7-point scale used in earlier experiments.

3.2 Results

We predicted that intention attributions would be higher in the high willpower condition. Our prediction and analysis plan were pre-registered.

There was a numerical difference in the predicted direction for mean intention attribution, but an independent samples t-test revealed that it was statistically insignificant (low/high: M = 4.93 / 5.15, SD = 1.71 / 1.7), t(207.99) = -0.93, p = .176 (one-tailed), d = -0.13 (see Fig. 3). One sample t-tests revealed that mean intention attribution was significantly above the neutral midpoint (=4) in both the low condition, t(104) = 5.61, p < .001, and the high condition, t(104) = 6.96, p < .001. Median intention attribution was 6 in the high condition, and it was 5 in the low condition. An exploratory Wilcoxon rank sum test failed to find a significant difference in median attribution, W = 5011.5, p = .121 (one-tailed).

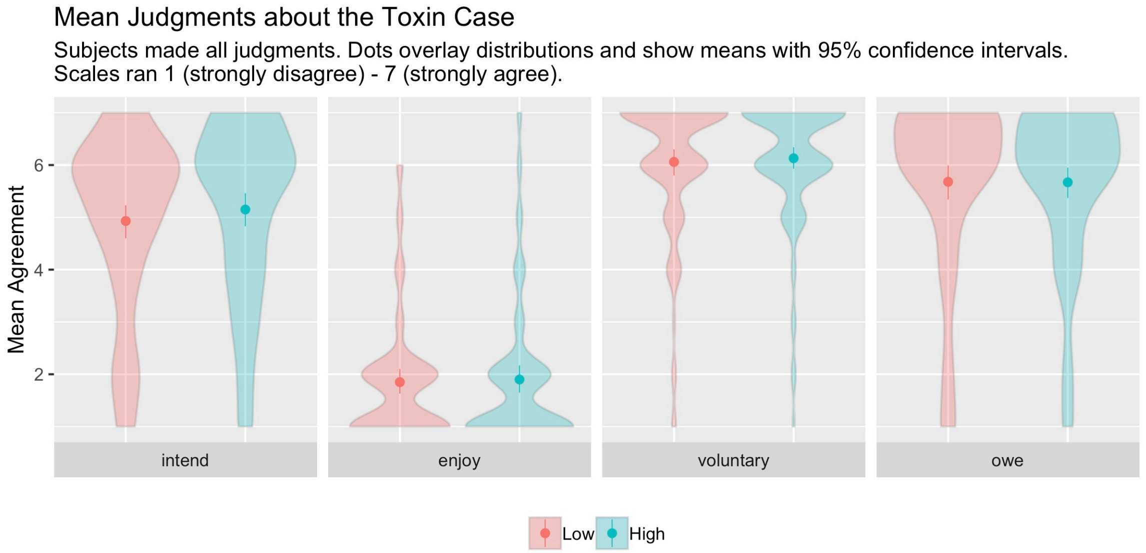

Fig 3: Experiment 3. Mean response overlaying distributions for the test statements (within-subjects) across two willpower conditions (low, high) (between-subjects). Scales ran 1 (“strongly disagree”) – 7 (“strongly agree”). Error bars show 95% bootstrapped confidence intervals.

3.3 Discussion

This experiment examined whether intention attribution in a classic toxin case was higher when the agent was described as strong willed compared to weak willed. As predicted, intention attribution was higher when the agent was strong willed, as measured by mean and median response, but the numerical difference did not reach statistical significance. Accordingly, the present results provide at best weak evidence for pure voluntarism. A possible explanation for this outcome is that intention is viewed as so voluntary that even very weak-willed people can form an intention with little difficulty.

4 Conclusion

We addressed two related conceptual questions that arise in the context of the toxin puzzle. On the one hand, is intention conceptually voluntary in the way that decision-making is assumed to be (i.e. is pure voluntarism about intention true)? On the other hand, is it conceptually possible to intend to do something despite lacking good reason to do it (i.e. is there a reason constraint)? Both of these research questions pertain to what is possible according to the ordinary, shared concept of intention, familiar to all of us from folk psychology. Our findings support the conclusion that pure voluntarism is true: on the ordinary view, intentions are viewed as voluntary. Our findings also support the conclusion that the reason constraint is false: it is conceptually possible to intend to do something despite lacking good reason to do it. Finally, following up further on the voluntariness of intention, we also looked for evidence that perceived willpower affects intention attribution. In line with our prediction, being described as strong-willed rather than weak-willed was associated with a numerical increase in people’s confidence that an agent formed an intention by choice, but this numerical difference did not reach the level of statistical significance. This could be because intention is viewed as being so deeply voluntary that not even a substantial lack of willpower can strip it of voluntariness.

Acknowledgments — For helpful feedback, we thank Carolyn Buckwalter, Al Mele, David Rose, and Angelo Turri. This research was supported by the Canada Research Chairs Program and the Social Sciences and Humanities Research Council of Canada.

References

Buckwalter, W., Rose, D., & Turri, J. (*in press*). Impossible Intentions. American Philosophical Quarterly.

Kavka, G. S. (1983). The toxin puzzle. *Analysis*, 43(1), 33-36.

Mele, A. R. (1992). Intentions, reasons, and beliefs: Morals of the toxin puzzle. *Philosophical Studies*, 68(2), 171-194.

Turri, J., Rose, D., & Buckwalter, W. (2018). Choosing and refusing: doxastic voluntarism and folk psychology. *Philosophical Studies*, 175(10), 2507-2537.