On the Robustness of Intuitions in the two best-known Trolley Dilemmas

Alex Wiegmann (awiegma@gwdg.de)
Institute of Psychology, Gosslerstraße 14
37073 Göttingen, Germany

Matthias Lippold (matthias.lippold1@stud.uni-goettingen.de)
Institute of Psychology, Gosslerstraße 14
37073 Göttingen, Germany

Robert Grigull (robert.grigull@stud.uni-goettingen.de)
Institute of Psychology, Gosslerstraße 14
37073 Göttingen, Germany

Abstract

The *Bridge* dilemma (pushing a heavy man from a bridge in front of a train that would otherwise kill five persons) and the *Switch* dilemma (redirecting a train that would otherwise kill five persons onto another track where it kills one person) are presumably the two best-known moral dilemmas in philosophy and psychology. In this paper we claim that people's intuitions about what to do in Bridge are robust, while intuitions about Switch can be influenced rather easily. In doing so, we strongly disagree with Broeders and colleagues (2011) who recently argued for exactly the opposite claim. We discuss their interpretation of previous findings that were supposed to motivate their claim, present findings from previous studies that strongly support my claim, and report on failed attempts to replicate and present an experiment in which participants were willing to revise their judgment for Switch but not for Bridge.

Keywords: moral judgment; trolley dilemmas; robustness of moral intuitions; priming; transfer effects.

Introduction

Bridge and *Switch* are presumably the two best-known hypothetical moral dilemmas. They were first extensively used as thought experiments in moral philosophy and later also in empirical studies in moral psychology (cf. Waldmann, Nagel, & Wiegmann, 2012; Gräfenhain & Wiegmann, 2012).

In both scenarios five people are threatened by an out of control train. In Bridge the only possibility to save the five persons is to throw a heavy person from a bridge in front of the train, resulting in killing the heavy person and saving the five (Thompson, 1985). In Switch the threatening train can be redirected away from the five onto another track where one person would die in the collision with the train (Foot, 1967). Research in moral psychology has shown that the majority of people disapprove intervening in Bridge while they tend to approve the action in Switch (Waldmann et al., 2012).

In their recent paper, Broeders, Bos, Müller, and Ham (2011) make extensive use of these two dilemmas. They argue that previous research, especially the research by Greene and colleagues (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004) indicates that people's decision in Switch is made fast and without hesitation, while it takes them longer and they are more hesitate to make a decision in Bridge. Following this line of argument, Broeders and colleagues (2011) claim that people's judgments in Bridge can easily be manipulated by priming them with rules as "save lives" and "do not kill", while this kind of priming supposedly has no effect on people's judgment in Switch. In three experiments they seemingly confirm this claim.

In this paper we argue for an opposite claim: Judgments concerning Switch can be manipulated rather easily while judgments concerning Bridge are rather robust.

Arguing against Broeders and colleagues' interpretation of previous findings

Broeders and colleagues' (2011) claim is motivated by the following line of argument. Research by Greene and colleagues (Greene et al., 2001, Greene et al., 2004) suggests that when people have to deal with Bridge the anterior cingulated cortex (ACC) shows increased activity. Activation of the ACC is assumed to indicate people's feeling of uncertainty. Moreover, people's longer reaction times in Bridge, as compared to Switch, are also assumed to indicate uncertainty. This uncertainty is then interpreted as people's struggling to choose between the two rules "Do not kill" and "Save Lives". Hence, by priming one of the rules and thereby making it more accessible to subjects, their intuition about what to do in Bridge allegedly follows the primed rule. In contrast to Bridge, Switch elicits low ACC activity and people respond fast to it, supposedly indicating certainty. Hence, judgments concerning Switch are assumed to be robust and not to follow the primed rule.

At first glance, this line of reasoning sounds plausible. However, a closer look at the cited studies reveals that they do not provide compelling evidence in support of Broeders' and colleagues' (2011) claim that people are uncertain of what to do in Bridge. Remember that this claim is based on two observations, namely people's longer reaction times and higher ACC activation in Bridge as opposed to Switch. However, there is no evidence that people's reaction-times

were longer for Bridge. While in their first fMRI study, Greene and colleagues (2001) did not report reaction times for Bridge, they explicitly state in their follow up study (Greene et al., 2004) that reaction times for Bridge were short.

What about the other finding that was also interpreted as people feeling uncertain about what to do in Bridge, namely the high ACC activation when people respond to this dilemma? First of all, the reported results in the fMRI-studies by Greene and his colleagues (2001, 2004) are based on brain activity averages for groups of dilemmas (about twenty in each group). Hence, to inferring conclusions from these averages to specific cases are just not valid.

Secondly, Greene and colleagues (2004) do not interpret high ACC activation as indicating uncertainty but as a conflict of emotion and cognition or, more precisely, cognitive effort to override a prepotent emotional response (cf. Stroop effect, Stroop, 1935). Their interpretation explains the aforementioned finding of people's longer reaction-times when choosing an utilitarian (cognitive) option in personal moral dilemmas, because people have to override a strong emotional response not to intervene (Greene et al., 2001). In the same way, Greene's et al. interpretation of ACC activity as indicating a conflict of emotional and cognitive (utilitarian) considerations can explain why reaction times were only longer for people under cognitive load, which were namely those who chose the utilitarian option (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008). In contrast, interpreting ACC activity as indicating uncertainty cannot account for these findings because it would predict longer reaction times in dilemmas with high ACC activity independently of the option (cognitive vs. emotional) people choose.

Thirdly, it is simply not the case that the cited studies provide any evidence for higher ACC activation in Bridge, as compared to Switch. In their first fMRI study (Greene et al., 2001) ACC activation is not measured at all. In the follow-up study (Greene et al., 2004) ACC activity in difficult and easy personal moral dilemmas was compared. Since Switch was in neither of these groups, we do not have any evidence on the level of ACC activation in this dilemma. Moreover, due to relatively fast reaction times, Bridge was classified as an easy personal dilemma and it was found that ACC activity in these dilemmas was significantly lower than in difficult moral dilemmas.

Previous research indicating intuitions about Bridge to be robust - but not about Switch

So far, we have dismissed Broeders and colleagues' (2011) argument that was supposed to motivate their claim that Bridge is easy to influence, as compared to Switch. Now we shall present empirical findings that strongly speak in favor of my claim, that is, if any of the two dilemmas can be influenced rather easily then Switch is the one.

Lanteri, Chelini, and Rizello (2008) presented participants with the Switch and the Bridge dilemma. In one condition, participants first had to judge Switch and then Bridge afterwards. In the other condition, Switch was preceded by Bridge. Although responses to Bridge remained unaffected by the order of presentation, fewer participants were willing to intervene in Switch when Switch was preceded by Bridge. The authors interpret their results as evidence that Switch may be perceived in more than one way, while they speculate that the emotions triggered by Bridge may be evolutionarily sound and hard wired into our species, making it more robust than reactions to Switch that are assumed to be a result of moral reasoning.

Lombrozo (2009) conducted a very similar experiment. The only difference to Lanteri et al. (2008) was that participants were allowed to read both dilemmas before they were asked to judge them. Again, participants who saw Switch first provided higher permissibility ratings than those who saw it after Bridge. Responses for Bridge were unaffected.

Petrinovich and O'Neill (1996) conducted several experiments in which participants were asked to judge a sequence of moral dilemmas, among them Switch and Bridge, where the order of presentation was manipulated between subjects. While ratings for Switch often significantly differed as a function of whether it was presented as the first or last dilemma, ratings for Bridge remained unaffected.

Finally, Wiegmann, Okan, and Nagel (2012) also found that people's judgments for Switch can be influenced by first presenting other scenarios, while people's judgments for Bridge were not affected. Moreover, Wiegmann, and Okan (2012) tried and failed to raise ratings in favor of the proposed action in Bridge. In one experiment they urged participants to justify their ratings in Switch, assuming that subjects' justification is something like "save as many lives as possible" and that this forced justification would raise subjects' ratings for Bridge. In another experiment, they tried to raise subjects' ratings for Bridge by first presenting them with a scenario in which there was only enough time to pull one of two switches. One switch prevented one person, the other three persons from being killed. Presenting this scenario first was also supposed to make a rule like "save the most lives possible" salient. However, neither attempt succeeded in influencing ratings for Bridge.

Replication Experiments

What follows are two attempts to replicate Broeders and colleagues' (2011) findings of their first experiment. We limit my replication attempts to their first of the total of three experiments for the following reason. All three experiments are based on the same rationale, namely to prime participants with one of the two rules. The only way the three experiments differ is how priming was implemented. In their first experiment, priming was implemented by asking participants to read a story and to answer two questions about the rule "Save lives" or "Do not kill". In Experiment 2 participants were asked to solve a sliding puzzle that resulted in a symbol supposed to prime participants with one of the two rules. In the third

experiment participants were subliminally primed. Hence, their first experiment is very similar to the experiments described in the preceding section. It might be possible, if unlikely, that the findings in their second and third experiment can be replicated even if it is not possible for the findings in their first experiment. However, since priming in their second and third experiment was implemented in a rather subtle way, as compared to reading a story in the first experiment, failing to replicate the findings in the first experiment would already strongly limit the scope of the claim that intuitions about Bridge can rather easily be manipulated while intuitions about Switch are rather robust.

In the light of what has been said so far, what prediction is to be made regarding Broeders and colleagues' (2011) experiment in which participants had to read stories designed to prime them with the rule "Save lives" vs. "Do not kill"? Surely, everything points to the prediction that Bridge will not be affected by their manipulation. With regards to Switch things are not that clear, because there are no previous experiments in which it was tried to influence ratings for Switch by priming rules.

First Replication Attempt

Participants 352 subjects, each receiving £ 0.50, were recruited via an online database located in the U.K. They were invited via an email. The email contained a link that directed them to the experiment. Mean age of the participants was 47 years and 4 months (*SD*=15 years, 7 months), 61% were female.

Design, Procedure, and Materials Participants were randomly assigned to one of the conditions of a 2 (primed rule: "Save lives" vs. "Do not kill") × 2 (dilemma: Bridge vs. Switch) factorial design.

After reading a cover story participants were asked to read a short story which was supposed to prime them with either the "save lives" or "do not kill"-rule. The "save lives" story goes as follows:

> During missions abroad soldiers have to apply to certain rules. This is called a mandate. During a conflict in 1994 in Congo, Africa, the mandate of the UN peace keeping force present in the country was to use armed violence to end the conflict. As a result of several warnings beforehand and during the conflict by the UN commander in chief on the spot, the Canadian lieutenant–general Roméo Dallaire, UN-soldiers were allowed to conduct offensive actions. They were both allowed to use their weapons to protect the civilian population, as well as when they themselves were attacked. The mandate was based on the principle "Save lives".

For the "Do not kill" version participants read:

> During missions abroad soldiers have to apply to certain rules. This is called a mandate. During a conflict in 1994 in Congo, Africa, the mandate of the UN peace keeping force present in the country was not to use any armed violence to end the conflict. Despite several warnings beforehand and during the conflict by the UN-commander in chief on the spot, the Canadian lieutenant–general Roméo Dallaire, UN-soldiers were not allowed to conduct any offensive actions. They were only allowed to use their weapons when they themselves were attacked. The mandate was based on the principle "Do not kill".

Afterwards, they were first asked to write down how they would act in line with the corresponding rule. Furthermore, they were asked to write down what the specific rule meant for them in their daily lives. Subsequently, participants were asked to rate how they feel at the moment on the 20 items of the PANAS, an instrument to collect self-report measures of emotion (Watson, Clark, & Tellegen, 1988). The PANAS was included as a filler task, as well as to determine whether the primes triggered any unwanted positive or negative affective reaction.

The PANAS was followed by asking participants to read and respond to either the Bridge or the Switch dilemma. Bridge reads as follows:

> You are standing on a footbridge crossing a railroad track. While you are standing next to a stranger, suddenly a runaway trolley comes hurtling down the railroad track. Further down the railroad track five people are working and they cannot possibly leave the railroad track in time. If the trolley proceeds on its present course it will crash into the five railroad workers and they will be killed in a fatal accident. The only way to save the five people is to push this man off the bridge and into the path of the trolley. The body of this person will break the speed of the trolley as a result of this the trolley will stop and the five persons will survive. The person thrown from the footbridge will certainly die.

Switch reads:

> You are standing beside the switch of a railroad track. Suddenly a runaway trolley comes hurtling down the tracks. Further-on down the railroad track five people are working and they cannot possibly leave the railroad track in time. If the trolley proceeds on its present course it will crash into the five railroad workers and they will be killed in a fatal accident. You can save these five people by diverting the trolley onto a different set of railroad tracks. The different railroad track has only one person on it, into which the trolley will crash. This person will be killed as a result of this.

After reading the dilemma participants were asked eleven questions about their willingness to intervene which they could indicate on a scale ranging from 1 (certainly not) to 7 (certainly yes). All items were then averaged to form a reliable scale indicating the willingness to intervene in the dilemma (α=.80).

Finally, participants were asked four questions to find out whether they were aware of the purpose of the experiment.

Results and Discussion Eleven subjects were excluded because at least one of two independent raters coded them as being aware of the purpose of the experiment.

As in the study by Broeders and colleagues (2011), the prime did not have an effect on the positive or negative subscale of the PANAS.

Figure 1 clearly shows that neither Bridge nor Switch was affected by a priming scenario. A 2*2 ANOVA yielded the typical main effect of dilemma, $F(1, 337)=80.70$, $p<.000001$, $\eta_p^2=.19$. However, there was no main effect of prime ($p>.75$) and no significant interaction ($p>.6$). Hence, Broeders and colleagues' finding could not be replicated although many more subjects participated in this experiment, resulting in a higher test power

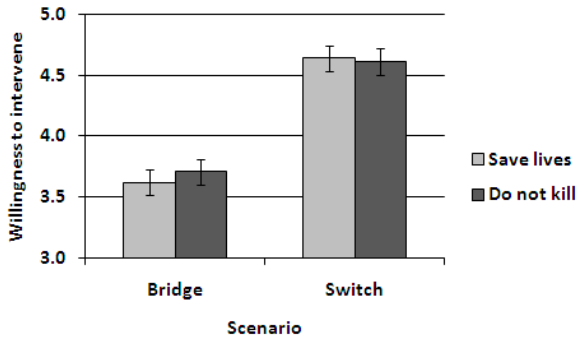

Figure 1: Willingness to intervene (on a scale from 1 to 7) in Bridge and Switch as a function of manipulated accessibility of the rules "Save lives" and "Do not kill". Higher bars indicate greater willingness to intervene. Error bars represent standard error of means.

Second Replication Attempt

This time we tried to replicate Broeders' et al. (2011) findings in our experimental lab in Goettingen. This was done to counter objections claiming that online experiments are not reliable (although the typical main effect of dilemma was found). The design and procedure was the same as in the first replication attempt with two exceptions. The PANAS was left out to strengthen the influence of the primes, and participants were only asked two questions concerning their willingness to intervene in Switch or Bridge since the correlation of the eleven questions asked in Broeders and colleagues' and in my first replication attempt was very high.

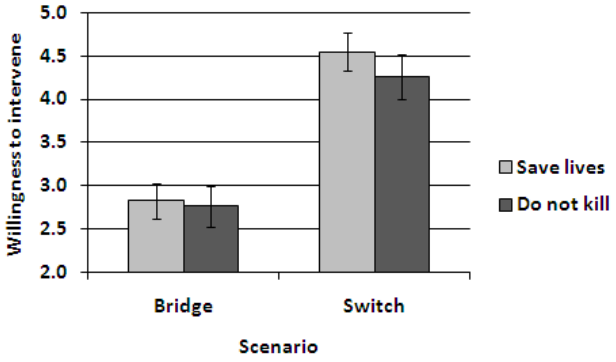

Figure 2: Willingness to intervene (on a scale from 1 to 7) in Bridge and Switch as a function of manipulated accessibility of the rules "Save lives" and "Do not kill". Higher bars indicate greater willingness to intervene. Error bars represent standard error of means.

Participants 220 participants, mostly psychology students, were recruited via the institute's database. Participants were credited with course credit or paid 7€/hour. Mean age of the remaining $N = 172$ participants was 24 years and 7 months (SD=6 years, 6 months), 77% were female.

Results and Discussion 48 participants were excluded from the analysis because they knew the dilemmas (41) or seemed to identify the purpose of the experiment. Figure 2 clearly shows that neither Bridge nor Switch was affected by a priming scenario. A 2*2 ANOVA yielded the typical main effect of dilemma, $F(1, 168)=47.55$, $p<.001$, $\eta_p^2=.22$. However, there was again no main effect of prime ($p>.45$) and no significant interaction ($p>.6$).

Judgment Revision Experiment

This experiment aims to investigate the robustness of Switch and Bridge by giving participants the chance to later revise their initial judgment.

Participants 158 subjects, each receiving £ 0.50, were recruited via an online database located in the U.K.

Design, Procedure, and Materials The experiment was conducted on the Internet. Upon clicking on a link they received via e-mail, participants were redirected to a website containing the experiment. They read general instructions familiarizing them with the rating scale and asking them to read the following scenario carefully and to take their task seriously. Afterwards, they were randomly assigned to one of two conditions. In Bridge_Switch participants were first presented with Bridge and then Switch, in Switch_Bridge ist was the other way around. In both conditions participants had the chance to revise their judgment for the first scenario after they had seen the second scenario. Both scenario descriptions were accompanied by an illustration of the initial situation.

For each scenario participants were asked whether the proposed action should be done. To indicate their judgment participants could mark one point on a 6-point likert scale ranging from 1 ("certainly no") to 6 ("certainly yes").

After participants were given the chance to revise their judgment for the first scenario, they were asked some demographic questions and a simple logical question to identify participants who did not take the experiment seriously.

Results and Discussion 27 participants were excluded from the analysis because they did not finish the experiment, finished it in less than 40 seconds, or failed to answer the logical question. As it can easily be seen in Figure 3 the aforementioned asymmetrical transfer effect between Bridge and Switch was replicated. While the ratings for Bridge did not differ significantly depending on whether it was

presented first (M=2.4, SD=1.51) or second (M=2.60, SD=1.64), $t(129)$=.74, p=.46, ratings for Switch were significantly decreased when Switch was presented second (M=3.02, SD=1.50), as compared to ratings for Switch when presented first (M=4.45, SD=1.36), $t(129)$=4.18, p<.00001.

When we consider the difference of a scenario's first rating vs. the revision rating a similar picture arises. The first rating for Bridge (M=2.4, SD=1.51) did not significantly differ from the revision rating (M=2.44, SD=1.59), t=.35, p=.73. In contrast to this, the revision rating for Switch (M=3.88, SD=1.63) did significantly differ from the first rating for Switch (M=4.45, SD=1.36), t=4.32, p<.0001.

The results strongly suggest that people's intuitions about Bridge are robust while their intuitions about Switch were significantly influenced when Switch was preceded by Bridge or when Bridge was presented after Switch and people were then given the chance to revise their judgment for Switch.

This pattern of results contradicts Broeder's et al. (2011) claim that people's intuitions about Switch are robust but not for Bridge.

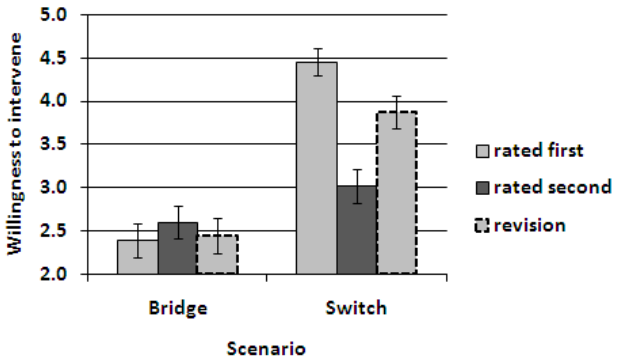

Figure 3: Willingness to intervene (on a scale from 1 to 6) in Bridge and Switch as a function of whether the scenario was shown first or preceded by the other scenario (first vs. second rating). The dashed lines represent revised ratings for the first scenario after participants were presented with the second scenario. Higher bars indicate greater willingness to intervene. Error bars represent standard error of means.

Conclusion

In this paper we argued that people's intuitions about Bridge are rather robust while their intuitions about Switch are rather easy to influence. This claim stands in sharp contrast to Broeders and colleagues' (2011) claims. We argued that Broeders and colleagues' interpretation of previous findings that were supposed to motivate claim is not sound. Moreover, we reviewed previous findings that strongly point in the opposite direction.

In line with my claim, replicating the findings of Broeder's et al. first experiment failed online as well as in the lab. Furthermore, the results the revision experiments also count in my favor.

Given the important role that Bridge and Switch play in philosophy as well as in psychology, it is important that wrong claims about them are swiftly corrected to avoid that new research is based on false premises.

Funding

This research was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG WA 621/21-1), and the Courant Research Centre 'Evolution of Social Behaviour', University of Göttingen (funded by the German Initiative of Excellence).

Acknowledgements

Thanks are due to Jonas Nagel, Michael Waldmann, Gunda Johannes, Matthias Lippold, and Robert Grigull for very helpful comments and proofreading.

References

Broeders, R., van den Bos, K., Müller, P. A., & Ham, J. (2011). Should I save or should I not kill? How people solve moral dilemmas depends on which rule is most accessible. *Journal of Experimental Social Psychology*, *47*(5), 923–934.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. O*xford Review*, *5*, 5–15.

Gräfenhain, M., & Wiegmann, A. (2012). The Scientific Study of Morals. In: Lütge, C. (Ed.), Handbook of the Philosophical Foundations of Business Ethics (chapter 81). Springer Press.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*, 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI study of emotional engagement in moral judgment. *Science*, *293*, 2105–2108.

Lanteri, A., Chelini, C. & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics*, *83*, 789-804.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, *33*, 273-286.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, *17*, 145-171.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*. *12*, 643–662.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison

(Eds.), *The Oxford Handbook of Thinking and Reasoning (pp. 364-389)*. New York: Oxford University Press.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063.

Wiegmann, A., & Okan, J. (2012). Order effects in moral judgments. Searching for an explanation. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1143-1148). Austin, TX: Cognitive Science Society.

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology, 25,* 813–836.