BAYESIAN ANALYSIS OF CHILDREN’S MORAL JUDGMENTS ON THE
TROLLEY CASES

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Paul C. Stey

________________________________________

Daniel K. Lapsley, Director

Graduate Program in Psychology
Notre Dame, Indiana
April 2014
BAYESIAN ANALYSIS OF CHILDREN’S MORAL JUDGMENTS ON THE TROLLEY CASES

Abstract

by

Paul C. Stey

The current project tested two competing views—with particular emphasis on the extent to which these competing accounts could explain an interesting and robust pattern of findings concerning the classic “trolley problem”. Specifically, this project tested aspects of Greene’s dual-process model (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001) and Mikhail’s (2000) theory of universal moral grammar. We also extended the research of Pellizzoni, Siegal, and Surian (2010), who utilized a sample of young children in an attempt to test Greene’s theory that moral judgments are driven by differences in the directness of harm. The current study found little evidence suggesting that children between the ages of 3 and 5 consistently aligned their moral judgments with the contact principle. Furthermore, there was also little evidence that children in this age group conformed their judgments to the doctrine of double effect, which would have supported Mikhail’s account. Finally, we also investigated candidate predictors of
the tendency to conform to these heuristic principles. This revealed that
inhibitory control negatively predicted the tendency to align one’s judgments
with the contact principle. This seems to support Greene’s theoretical view. In
addition, we found that participant age was a negative predictor of the tendency
to align one’s judgments with the doctrine of double effect—an unexpected
result. This project adopted a Bayesian approach because of its overall flexibility
and the ease with which missing data can be handled.
pour ma famille
# TABLE OF CONTENTS

LIST OF FIGURES ................................................................. iv

LIST OF TABLES ....................................................................... v

ACKNOWLEDGMENTS ............................................................... vii

INTRODUCTION ...................................................................... 1

PRESENT STUDY ...................................................................... 20

METHOD ............................................................................... 27

RESULTS ............................................................................... 42

DISCUSSION ......................................................................... 70

REFERENCES ......................................................................... 78

APPENDIX A .......................................................................... 83

APPENDIX B .......................................................................... 95
LIST OF FIGURES

Figure 1. Figure 1a (left panel). The Loop Track case in which the protagonist can flip a switch to divert the trolley from the main track to the side track. This would kill the lone bystander, but save the five on the main track. Figure 1b (right panel). The Loop Weight case in which the protagonist can flip a switch to divert the runaway trolley from the main track to the looping side track. This would kill the lone bystander on side track, but save the five on the main track. This differs from the original Loop Track case in that the lone bystander’s death is merely a foreseen side effect in this version, and not a necessary means to saving the five as it is in the Loop Track case ..........................................................9

Figure 2. Illustration of Footbridge case of trolley problem. The protagonist must use direct physical contact to cause harm, and the harm is the means by which the five men are saved .................................................................13

Figure 3. Illustration of Bystander case. The protagonist does not use direct physical contact to cause harm, and the harm to the bystander is not the means by which the five men are saved .................................................14

Figure 4. Figure 4. Summary of participants’ ages in months. The distribution is nearly uniform, but lacks participants toward the very high end of 70 months of age ...........................................................................................................28

Figure 5. Illustration of Bystander_absent case in which the protagonist can flip the switch to divert the trolley to the empty side track. This vignette serves as a positive control, where the rational response is to intervene ........31

Figure 6. Illustration of Bystander_inv case in which the protagonist can flip the switch to divert the trolley to the side track, which would sacrifice 5 people to save 1 person. This vignette serves as a negative control, where the rational response is non-intervention ........................................32
Figure 7. Summary of missing data in $N \times p$ matrix of variable sorted by age in months. Red blocks indicate missing observations. Black, white, and grayscale blocks represent observed data, with higher values on a variable are denoted by darker shades. This figure is sorted by the variable $\text{age}_m$, such that younger participants' records appear lower on figure ..........................................................44

Figure 8. Posterior distribution of $\theta_{\text{contact}}$ suggesting that the probability of conforming to the contact principle is not different than what would be expected with participants responding at chance ....................................47

Figure 9. Posterior distribution of $\theta_{\text{dfe}}$ suggesting that the probability of conforming to the doctrine of double effect is less than what would be expected with participants responding at chance ....................................49

Figure 10. Summary of trace plots from model M1a above. These illustrate fast mixing of chains, and overall model convergence ........................................54

Figure 11. Posterior distribution of the effect of inhibitory control on the tendency to endorse intervention in the footbridge case .........................58

Figure 12. Trace plots for parameter estimates in M2a. These suggest fast mixing of Markov chains, and that the model converged appropriately ..........60

Figure 13. Trace plots of Markov chains for model M3a. These illustrate fast mixing and appropriate convergence .................................................64

Figure 14. Summary of trace plot for Markov chains used to estimate parameters in M4a. These illustrate the model converged appropriately ..........68

Figure 15. Image used for PAL receptive vocabulary task. Participants were asked to point to the picture of the object that was named ...............93
LIST OF TABLES

Table 1. Sample Ordering of Tasks ........................................................................................................40

Table 2. Inhibitory Control and Effortful Control Predicting Conforming to Contact Principle .................................................................53

Table 3. Models Predicting Tendency to Conform to Contact Principle ..........56

Table 4. Priors and Link Functions Used in Sensitivity Analyses ..................56

Table 5. Age Predicting Conforming to Doctrine of Double Effect .................59

Table 6. Models with Age Predicting Doctrine of Double Effect .................61

Table 7. ToM Predicting Conforming to Doctrine of Double Effect ............63

Table 8. Models with ToM Predicting Doctrine of Double Effect .................65

Table 9. Model for Age and ToM Interaction Predicting Conforming to Contact Principle .....................................................................................67

Table 10. Models for the Interaction of Age and ToM Predicting Doctrine of Double Effect .....................................................................................69
AWKNOWLEDGMENTS

By almost any metric one would choose to apply, I have been exceptionally lucky in my life. Of particular importance to this project is that I have been lucky to be surrounded with bright and generous individuals. I am thankful for the chance to acknowledge them here.

First and foremost, I am grateful beyond words to my advisor and mentor Dan Lapsley. I am thankful for his support, encouragement, and his faith in me these past five years. It would be difficult to place a value on the experience that I have gained working in his lab, but I know that I owe him more than I can ever repay.

I am also hugely indebted to Ying (Alison) Cheng. I owe Alison not only for her guidance and thoughtful comments on this project, but also for her help and encouragement in pursuing quantitative interests. Having Alison as a professor, and then a committee member, proved to be a very fortunate series of events in my life.

I also owe a great deal to the other members of my committee. Nicole McNeil and Jill Lany were helpful throughout the course of this project, but I
am especially grateful for their generosity and kindness in helping me with refining the tasks and with the recruitment of participants. Participant recruitment is always a challenge, and I cannot imagine how I would have gotten by without the profound generosity of both Jill and Nicole. I am similarly indebted to Nicole and Jill’s lab managers, Julie Johnson and Olivia Wen. Both were enormously helpful and kind; I have trouble imagining where I would be without Nicole, Jill, Julie, and Olivia.

Many others helped with recruitment. Amy Nuttall was especially helpful when I first launched this project, and I am grateful for that. But I am also grateful for her kindness and support throughout the entire project. I am grateful to Kristin Valentino for the same reason; her help proved invaluable at a time when I desperately need it. Terri Kosik also helped me more than I can say. If not for her generosity, I would still be collecting data instead of finalizing my document. Andrea Christensen deserves a special mention, too. I am supremely grateful to her and her husband, Cooper, not only for their willingness to spread the word to potential participants, but also for the years of kindness, generosity, and support they have shown.

Everyone at the Notre Dame Center for Children and Families also deserves an enormous amount of credit. I am especially thankful for the consideration of Julie Braungart-Rieker and Jen Burke LeFever. I cannot
overemphasize the extent to which their generosity made my life easier. I am also hugely appreciative of Terri Sweeney, whose helpfulness, kindness, and positivity made every day at the Center for Children and Families an absolute pleasure.

And on the subject to data collection, perhaps the biggest debt I owe is the one I owe to my experimenters: Jessie Kusina, Maddie Jarrett, Lisa Rague, and Claire Cosgrove. These four young ladies sacrificed a great deal to help me, and I suspect that they probably got more than they bargained for in the process. I hope that I can eventually repay the favor.

Although it should go without saying, I must also thank all the participants and their families for their willingness to be a part of this project. They helped me more than they realize, and I hope that at some point in their lives they are in a position to benefit similarly from the kindness of a stranger.

I am also hugely indebted to Sara Fulmer. Her support has meant everything, and I doubt that I would be at this point were it not for her. Sara is another prime example of the extent to which I have been exceptionally lucky in my life.

Finally, I am most grateful to my parents, George Stey and Marie-Madeleine Stey. We are all our parents’ children, and for me, this has been particularly fortuitous. I am certain that I have them to thank for all my intellectual curiosity, ability, and work ethic. And of course, I would not be here
without all the material sacrifices they made for my siblings and myself. All of my accomplishments belong to them; I owe them everything.
INTRODUCTION

In recent years, there has been a revitalized interest in moral psychology research. The revived interest is at least partly due to the increasingly interdisciplinary nature of this area of research. Within the field of psychology, there was a time when developmental researchers dominated this area. More recently, moral psychology has grown beyond these boundaries and is now studied by researchers from all areas of psychology and neuroscience. Moreover, researchers outside of psychology and neuroscience have also begun working on problems once regarded as strictly the domain of psychologists. A result of this renewed enthusiasm in moral psychology has been an abundance of new theories that attempt to explain one of the most fundamental questions in psychology: 

*How do we decide what is right or wrong?* The modern empirical study of this question has its roots in the Piagetian (1965) and Kohlbergian (1969) traditions, but the recent growth and the interdisciplinary nature of this area of research has yielded a number of insights as well as important new directions for research. The current project aims to shed light on this key question.
One recently developed framework comes from the research of Mikhail (2000, 2007a, 2011), a philosopher and legal scholar who draws heavily from the work of Chomsky (1956). Mikhail argues that some features of moral cognition are the product of a native faculty that takes in as inputs the causal and intentional features of a situation and returns a judgment based upon the application of certain rules (e.g., “Intentionally causing harm to someone is bad”). The specific content of these rules, Mikhail argues, amounts to the “moral grammar” that individuals acquire over the course of development within their culture. It is worth noting the parallels between Mikhail’s theory and that of Chomsky. Whereas Chomsky suggested that humans are equipped with a language acquisition device, Mikhail argues that humans are born with a morality acquisition device. Like the content of grammar, the specific content of one’s morality will be largely determined by the cultural surroundings during one’s early years. Thus, according to Mikhail, judgments of right and wrong are guided by a native system of “moral grammar”; and this system, though shaped by one’s cultural environment, shows remarkable consistency across cultures (Mikhail, 2000). As evidence of this, Mikhail points to the cross-cultural research showing near-universal prohibitions against certain acts (e.g., murder, rape, and incest). For this reason, Mikhail suggests that we possess a universal moral
grammar, which disposes us to learn and acquire moralities in a similar manner and to a similar extent that we acquire languages.

Mikhail’s (2000, 2007a, 2011) research on universal moral grammar utilizes various forms of the classic “trolley problem” (Foot, 1967; Thomson, 1976), which has been studied in philosophy for decades. The essential feature of all trolley cases is that the protagonist in the vignette is faced with a decision of whether or not to harm one person in order to save the lives of several people. These dilemmas have become a standard methodological tool in research on moral psychology (e.g., Broeders, van den Bos, Müller, & Ham, 2011; Cushman, Young, & Hauser, 2006; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Kamm, 1998; Stey, Lapsley, & McKeever, 2013). Mikhail points out that a consistent finding emerges throughout his studies: individuals are sensitive to the intentionality involved in causing harm, and also to the total acts of battery involved in causing the harm.

The “Personal” and “Impersonal” Distinction

Greene and colleagues (Greene et al., 2001) have been vocal about their disagreement with Mikhail’s view (e.g., Greene, 2007) and they suggest a different theory. They argue that our moral judgments are often driven by affective responses that we have towards direct physical harm. More specifically,
Greene suggests a dual process model whereby characteristically deontological judgments tend to be produced by affective responses, while characteristically utilitarian responses are the product of deliberate reasoning.

**Normative ethical theories.** The term “deontological judgments” refers to those moral judgments that emphasize individual’s rights and duties when attributing moral qualities to actions. Deontological ethics, which is most closely associated with the work of Kant (1785/1959), presupposes that certain actions are simply wrong by their very nature, and that it is not the extent to which they produce good outcomes that ought to determine whether an action is right or wrong (e.g., stealing is always wrong, regardless of good outcomes). Instead, actions are deemed right or wrong according to the intrinsic features of the action, and how these align with predetermined principles that we hold as the moral standard—for Kant, the moral standard was the categorical imperative.

Another school of thought within normative ethics is that of utilitarianism. This normative ethical theory is most closely associated with the work of Bentham (1789/1948) and Mill (1861/2001). According to a common formulation of utilitarianism, the outcomes of actions are the *only* thing that ought to be considered when determining the extent to which an action is right or wrong. For utilitarianism, the right action is the one that maximizes the balance of good over evil, pleasure over pain, or joy over suffering. The morally
right action is the one that leads to the best possible state of affairs for all those involved. Thus, the term “utilitarian judgment” refers to those judgments that give preference to the overall balance of good over evil (e.g., saving more lives when it is possible).

Greene’s dual process model asserts that when considering the moral qualities of an action that involves “personal” harm, we are primarily guided by affective processes that tend toward deontological ethics. But, when evaluating the moral qualities of an action that does not involve personal harm, we instead tend to make judgments using a more calculating and rational process that leans toward utilitarianism.

An example that Greene (2003) provides is rather illustrative of the distinction between personal and impersonal harm, as well as the role of deontology and utilitarianism. Imagine a man, call him “Jones”, driving his sports car along the winding roads of the California mountains. Jones comes across a hiker who has been injured and is bleeding severely. He is in need of immediate medical attention. However, Jones recently spent $200 having his car detailed. The injured hiker would certainly bleed on the recently detailed car. When considering this scenario, it seems obvious that the proper thing for Jones to do would be to help the injured hiker. It seems simply wrong for him not to do so for fear of ruining his recently detailed upholstery. But consider a separate
scenario now. In this scenario a man, call him “Smith”, is at home when two individuals working for the non-profit group *Feed the Children* visit him. These men ask Smith to donate $200 to help feed starving children in sub-Saharan Africa. In this case, it does not seem so obvious that Smith ought to oblige. Greene suggests that the tension between personal and impersonal harm explains our differing intuitions with respect to these cases.

**Evidence for the personal–impersonal distinction.** Greene and colleagues (Greene et al., 2001) provide supporting evidence for this account using multiple methods. In their 2001 paper Greene and his colleagues describe a study in which they presented participants with various ethical dilemmas, broadly inspired by the trolley problem. There were three sorts of dilemmas: (1) “personal” dilemmas, which involved a protagonist causing direct physical harm to the victim, (2) “impersonal” dilemmas, in which the protagonist causes harm to the victim through some indirect method (e.g., pulling a lever), and (3) non-moral dilemmas, in which the protagonist is faced with a decision between options—one of which leads to a sub-optimal outcome (e.g., choosing between a bus or a train when under a time constraint). Participants were presented with 60 dilemmas from these three categories (20 of each) while experimenters recorded brain activity using functional magnetic resonance imaging (fMRI). The findings indicated that when participants were responding to personal moral
dilemmas there was increased activity in the brain regions associated with the processing of emotional information, including the: medial frontal gyrus, posterior cingulate gyrus, and the angular gyri (Maddock, 1999; Reiman, Lane, Ahern, Schwartz, & Davidson, 1997). Additionally, response time data revealed a significant dilemma-type by judgment interaction. This seemed to indicate that when participants endorsed intervention in the personal harm dilemmas, it took them significantly longer than either the impersonal moral dilemmas, or the non-moral dilemmas. Greene and colleagues explain this interesting finding by suggesting that those who endorsed sacrificing one person to save the lives of many in the personal harm dilemmas had to overcome a prepotent affective response condemning intervention. Thus, these longer response times are a function of wrestling with—and eventually overriding—one’s intuitive response. This aligns with Greene’s dual process model of moral cognition, which views characteristically utilitarian judgments as driven by conscious decision-making processes.

**Criticisms of the personal–impersonal distinction.** The findings reported by Greene and colleagues (2001) have recently been called in to question. Specifically, McGuire and colleagues (McGuire, Langdon, Coltheart, and Mackenzie, 2009) criticize the work of Greene and colleagues for several methodological flaws, and allege that their findings amount to artifacts of these
flawed methods. McGuire and colleagues were given access to, and reanalyzed, the raw data reported in the paper by Greene and colleagues (2001). McGuire and colleagues examined the dilemmas individually. This revealed that a few of the personal dilemmas with very low percentages of participants judging the actions to be appropriate had strikingly fast response times (i.e., participants were very quick to reject something as impermissible). Once these “poorly endorsed” items were removed from the analyses, the crucial interaction effect was no longer apparent. Thus, rather than the longer response times for personal moral dilemmas that were judged appropriate driving the interaction effect (as Greene and colleagues had claimed), it seems instead that the very fast responses judging the dilemmas as inappropriate was generating the significant interaction.

Greene’s work has also been criticized by Mikhail (2007b) for its failure to account for an interesting and robust pattern of findings from the literature on trolley problems. This interesting finding emerges when contrasting two of the most well-known trolley cases: the Loop Track case (see Figure 1a below), and the Loop Weight case (Figure 1b). In the Loop Track case there is a runaway trolley car coming down a track. If left unimpeded, the trolley will hit five bystanders on the main track. However, the vignette’s protagonist is positioned at a switch that can divert the trolley from the main track to a looping side
track. The looping side track has one large man standing on the track who will be hit if the protagonist diverts the trolley.

Figure 1a (left panel). The Loop Track case in which the protagonist can flip a switch to divert the trolley from the main track to the side track. This would kill the lone bystander, but save the five on the main track. Figure 1b (right panel). The Loop Weight case in which the protagonist can flip a switch to divert the runaway trolley from the main track to the looping side track. This would kill the lone bystander on side track, but save the five on the main track. This differs from the original Loop Track case in that the lone bystander’s death is merely a foreseen side effect in this version, and not a necessary means to saving the five as it is in the Loop Track case.

When faced with Loop Track case, many participants are reluctant to endorse the protagonist’s intervention. But consider the nearly identical Loop Weight case. In this version of the trolley problem, the circumstances are virtually identical to the Loop Track case. Even a casual glance at Figures 1a and 1b reveals the only substantive difference between these two cases is the presence of a large weight on the looping side track. But as we see below, this subtle
difference turns out to be quite important in its impact on judgments of permissibility.

**The doctrine of double effect.** As Mikhail (2000) and others (e.g., Cushman et al., 2006; Stey et al., 2013) have shown, individuals are more inclined to endorse intervention in the Loop Weight case than they are in the original Loop Track case. Mikhail (2000, 2007b) explains this interesting finding by pointing out that the presence of the large weight in the loop track case changes the lone bystander’s death from a necessary causal event to a foreseen side effect. Reasoning from counterfactuals, it is clear that in the Loop Track case the death of the lone bystander on the looping side track is the necessary means by which the five on the main track are saved. Were it not for the presence of the lone bystander on the main track, the five would die regardless of the protagonist’s actions. The same is not true in the Loop Weight case. The presence of large weight on the looping side track means that the lone bystander’s death is merely an unfortunate consequence rather than the means by which the five lives are saved. This subtle yet critical difference illustrates the doctrine of double effect, a philosophical principle with a history going back to Aquinas (13th c./1988).

The doctrine of double effect suggests that it may be permissible to cause harm in order to bring about a greater good provided that the harm is not the
necessary means towards that greater good, but only a foreseen side effect. The *Loop Track* case is a good example of harming as the means towards achieving a greater good. In this dilemma, the lone bystander’s death is the means by which the five lives are saved. The *Loop Weight* case, however, is an example of when the harm is merely incidental and not the means by which the greater good is achieved.

Despite its inability to account for the pattern of findings illustrated by the classic *Loop Track* and *Loop Weight* cases, Greene’s two-process model of moral cognition (e.g., Greene et. al., 2001) still enjoys considerable support from researchers. In fact, the distinction between personal and impersonal harm figures prominently in a recent study that investigated moral judgments in 3- to 5-year-old children.

**The Personal–Impersonal Distinction in Children**

A paper by Pellizzoni, Siegal, and Surian (2010) investigated the extent to which young children made moral judgments that align with Greene’s personal–impersonal distinction. In three studies Pellizzoni and colleagues attempted to show that children between the ages of 3 and 5 were more inclined to make
utilitarian moral judgments (i.e., maximizing good outcomes) when doing so would not involve causing direct physical harm to a victim.

Pellizzoni and colleagues (2010) recruited children between 3 and 5 years of age. These children were presented with two different versions of the classic trolley problem: the *Footbridge* case, and the *Bystander* case (Foot, 1967; Thomson, 1976). In the *Footbridge* case the protagonist is standing on a footbridge that passes over train tracks (see Figure 2 below). He sees there is a runaway trolley coming down the tracks. There are 5 people standing on the tracks in the path of the trolley. They cannot hear the trolley and will not be able to get off the track in time. Standing next to the protagonist on the footbridge is a very large man wearing a heavy backpack. The protagonist knows that the only way to stop the trolley is to push the large man onto the path of the trolley car. The large man will die, but this will slow the trolley down giving the 5 people time to escape. Research suggests that the majority of individuals are not willing to endorse pushing the man in the *Footbridge* case (Mikhail, 2000, Cushman et al., 2006).
The *Bystander* case is similar, but contains a few crucial differences (see Figure 3 below). Again, there is a runaway trolley coming down the tracks toward 5 people. This time, the protagonist is beside the tracks standing next to a switch that will send the train down a side track. However, there is one man standing on the side track. If the protagonist flips the switch, the 5 people will be saved, but the man on the side track will die. Unlike in the *Footbridge* case, the majority of people are willing to endorse flipping the switch and sacrificing the one man (Mikhail, 2000, Cushman, et al., 2006).
Figure 3. Illustration of Bystander case. The protagonist does not use direct physical contact to cause harm, and the harm to the bystander is not the means by which the five men are saved.

There are a variety of explanations as to why individuals are willing to allow sacrificing the man in the Bystander case but not in the Footbridge case. The view advanced by Pellizzoni and colleagues is that of Greene and colleagues (2001), which suggests that the two cases differ with respect to their ability to elicit emotion. In particular, Greene and colleagues propose that the direct physical contact in the Footbridge case produces a negative affective response that is not present when participants consider the Bystander case. Greene and colleagues actually used this pair of dilemmas in their influential 2001 paper, which reported that the Footbridge case (and similar dilemmas involving
“personal force”) was more likely to activate brain regions that have been implicated in the processing of emotions (Maddock, 1999; Reiman et al., 1997).

**Cause for skepticism.** There are at least three reasons to be a bit wary about the conclusions of Pellizzoni and colleagues. These are outlined below. The first concerns a methodological fault, which may have lead the authors to erroneously implicate the directness of harm as driving the moral judgments. The second reason for skepticism concerns the authors’ assumption that participants tracked the causal chain in the vignettes appropriately. And the third is a failure to address the important results of a control scenario. We address these more thoroughly below.

*A methodological reason for doubt.* Similar to prior findings with adults (e.g., Cushman, et al., 2006; Greene et al., 2001; Mikhail, 2000), Pellizzoni and colleagues found that children in their study supported intervention in the *Bystander* case, but not in the *Footbridge* case. The authors concluded that children made judgments according to the contact principle (Cushman et al., 2006), which suggests that harm caused by physical contact is morally worse than equivalent harm caused without physical contact. This aligns with Greene’s account. However, this conclusion may be flawed, as the *Bystander* case and the *Footbridge* case differ with respect to both physical contact (i.e., the contact principle) and with respect to the doctrine of double effect.
Recall that according to the doctrine of double effect, it is permissible to cause harm in order to bring about a greater good provided that the harm is not the necessary means by which the greater good is achieved. In the *Footbridge* case, pushing the large man into the path of the trolley is the necessary means by which the 5 people are saved. This same is not true for the *Bystander* case. If we examine this case closely, it becomes apparent that the man on the side track plays no causal role in saving the 5 people; instead, his death is merely an unfortunate consequence. If the man were not on the side track in the *Bystander* case, the protagonist would still be wise to flip the switch to divert the trolley.

Thus, the *Bystander* case involves the protagonist inflicting harm without physical contact, and without the intention to cause harm, whereas in the *Footbridge* case the protagonist causes direct physical harm that is also intended. Because of this, the comparison of these two trolley cases cannot be said to test the extent to which participants conform their judgments to the contact principle. At best, the reasonable conclusion is that the children in the study by Pellizzoni and colleagues (2010) were reacting either to the difference in physical contact, or to the difference in intentionality, or to a combination of these factors.

Since the methods used by Pellizzoni and colleagues (2010) conflated the contact principle and the doctrine of double effect, we have cause to remain a bit
skeptical of their conclusions. At the very least, further research is warranted before we can affirm that the contact principle guides the moral judgments of children between 3 and 5 years of age. Furthermore, it might be the case that when we untangle the doctrine of double effect and the contact principle, we find that the former was the force driving the effect Pellizzoni and colleagues report.

*Theory of mind and the doctrine of double effect.* The possibility that the doctrine of double effect—and not the contact principle—produced the pattern of judgments observed by Pellizzoni and colleagues suggests another important consideration. Prior research (Cushman & Young, 2011) has shown that the tendency for individuals to make moral judgments that align with the doctrine of double effect is dependent on their sensitivity to the intentions of others. This sensitivity to the intentions of others is a facet of one's theory of mind (ToM).

Since the methodology used by Pellizzoni and colleagues (2010) conflated the contact principle with the doctrine of double effect, it is important to consider the extent to which ToM may have influenced moral judgments. This is especially important given the age range of the children in their sample.

Although there is controversy in the field about how best to assess ToM (e.g., Baillargeon, Scott, & He, 2010), a wealth of prior research suggests that ToM generally develops around 4 years of age (e.g., Perner, Leekam, & Wimmer, 1987;
Rubio-Fernández & Guerts, 2013). As it turns out, this age range is rather important in the findings of Pellizzoni and colleagues (2010).

**Another methodological reason for doubt.** Another reason to be a bit skeptical with respect to the conclusions of Pellizzoni and colleagues (2010) pertains to the manner in which the stimuli themselves were presented. In particular, for both the *Bystander* case and the *Footbridge* case, the stimuli were presented in such a way that it is unclear whether or not the participants actually represented the causal chain of events.

The presentation of stimuli involved using models with toy figures representing people. The experimenter then read a script to the participant describing the nature of the dilemma, and asked the test question. For example, in the *Footbridge* case, participants were asked: “What should John do? Push the person or not push the person? Point to the picture showing what John should do.” The participant pointed to one of two pictures illustrating the two possible outcomes. In the *Footbridge* case, the participant would have seen two pictures. In one the large man lay on the track ostensibly having been pushed by the

---

1 Although studies have found children who demonstrated competence with ToM prior to 4 years of age (e.g., Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007), these studies tend to rely on spontaneous response paradigms (e.g., violation-of-expectation, and anticipatory-looking). There is debate about the extent to which these spontaneous response paradigms are tapping the same construct as more traditional false-belief tasks (e.g., Wimmer & Perner, 1983) with explicit response formats. A discussion of the nuances of implicit versus explicit beliefs is beyond the scope of this project, but see the paper by Low and Perner (2010) for a more thorough discussion of the subject.
protagonist, thereby saving the five people on the main track. In the other, the large man remained on the footbridge with the protagonist, and the five people on the main track have been hit.

Essentially, the question instructing participants on how to respond assumes that participants who indicate the desired outcome will have accurately represented the causal chain that lead to the outcome. While lessening the task demands is a often a desirable goal, in this case there is reason to believe that doing so amounts to changing the nature of the research question. Instead of testing whether or not participants think action $A_1$ is preferable to action $A_2$, it seems that the authors actually tested whether outcome $O_1$ is preferred to outcome $O_2$. The authors’ conclusions, then, are contingent on whether or not we accept that when a participant indicates that $O_1$ is preferred they have reasoned that the outcome $O_1$ was brought about by the action $A_1$. That is, we must make the assumption that the participants were able to track the chain of causality.

Assume we are willing to grant that participants accurately represented the causal chain and understood the implications of their responses. What would have been required of them? First, they would have had to correctly deduce that the question being posed was about events in the future. The relevant clue to this is the use of the modal verb “should”, which modifies the verb “do” in the
question above. This is why adults have no difficulty understanding what is being asked. Second, participants would have had to take this information and make the determination that the two candidate outcomes represented in the two pictures did correspond to future events, and that these future events are mutually exclusive. Finally, participants would have had to connect the candidate futures with the two proposed courses of action (e.g., push the man, or do not push the man). This requires backward induction.

Whether or not the participants were able to make all these necessary connections is an empirical question. But assuming that the connections have been made—as Pellizzoni and colleagues do—seems a bit too charitable. This is particularly true given some evidence that children under 5 years of age have trouble understanding cause and effect (e.g., Bullock, 1984) and have difficulty grasping the role of temporal order in causation (e.g., McCormack & Hoerl, 2007). Consequently, we should be concerned with a method that assumes so much of young children’s abilities.

**Age and failed control dilemmas.** One additional reason to exercise caution in interpreting the findings of Pellizzoni and colleagues (2010) comes from some pilot data that seem to undermine their argument. In Experiment 3, the authors tested children’s inclination to endorse the protagonist’s intervention regardless of outcome that would result. Participants were presented with what
was essentially a control scenario—specifically, an “inverted” Bystander case, where 5 people would be sacrificed to save 1. This would test the alternative hypothesis that children were simply endorsing intervention without regard to the outcome that resulted. In pilot testing this study the authors found that:

...many 3-year-olds could not clearly attend to the story scenario. Instead they simply seemed to interpret the test question as a request to indicate that it was better to act to save others and to answer that action was required (Pellizzoni et al., 2010, p. 269).

These findings weaken the authors’ case, and are an indictment of the previously reported conclusion that children as young as 3 years of age tend to act in accordance with the contact principle. Unfortunately, the authors give no further attention to this point.
PRESENT STUDY

Thus, the goals of the current project were threefold. First, we retested the hypotheses of Pellizzoni and colleagues (2010) using a more precise methodology. Specifically, we attempted to disentangle the role of the contact principle and the doctrine of double effect in the stimuli, and we altered the response format such that it makes no assumptions about children’s ability to track the chain of causality backwards from the outcome. Second, in addition to employing a more subtle methodology to retest their hypotheses, we also examined the nature of age difference in participants’ patterns of judgment. This was particularly important given the failed control vignettes that Pellizzoni and colleagues report for the youngest participants in Study 3. Finally, we investigated the role of inhibitory control and ToM as predictors of the tendency to conform to the two heuristic principles under investigation.

Recall that Greene and colleagues (2001) believe conforming to the contact principle consists in relying on one’s prepotent and intuitive response. If they are correct, then we ought to expect that inhibition could be predictive of the tendency to conform to the contact principle—specifically, we would expect it to be a negative predictor. All else being equal, it is reasonable to expect that those
individuals with less inhibitory control would be more disposed towards conforming to the contact principle. This follows naturally from the understanding that conforming to the contact principle—for Greene—means deferring to one’s intuitive responses.

Additionally, we anticipated the role of ToM would be an especially important predictor of the tendency to conform to the doctrine of double effect. Our reasoning is based on prior research (e.g., Cushman & Young, 2011) has shown that the tendency to make these sorts of moral judgments is related to one’s sensitivity to the intentions of others—a component of ToM.

Thus, the present study tested two hypotheses concerned with the extent to which participants’ patterns of moral judgments aligned with two heuristic principles. Additionally, we tested the role of age as a predictor of participants’ patterns of ratings. We also tested hypotheses pertaining to the importance of inhibitory control and ToM as predictors of conforming to the contact principle and the doctrine of double effect, respectively. And finally, we tested a hypothesis proposing that the interaction between participants’ age and ToM is a significant predictor of their tendency to conform their judgments to the doctrine of double effect. The details are outlined below.
Hypotheses

**Hypothesis 1 (H1).** This first hypothesis re-addressed the question posed by Pellizzoni and colleagues (2010). Specifically, H1 asserted that participants would tend to make moral judgments that conform to the contact principle. That is, they would tend to make utilitarian moral judgments in cases where doing so does not involve causing harm to a victim through direct physical contact. If participants were conforming their judgments to the contact principle, we would expect them to more readily approve of intervention in the *Loop Track* case compared to the *Footbridge* case.

**Hypothesis 2 (H2).** H2 supposed that participants would be inclined to conform their pattern of judgments to the doctrine of double effect. That is, they would tend to prefer harm that brings about a greater good provided that the harm is not the means by which the greater good is achieved. If participants were indeed conforming their judgments to the doctrine of double effect, we expected that they would more readily endorse intervention in the *Bystander* case and not in the *Loop Track* case. Note that though H1 and H2 are not in opposition, they can be thought of as competing accounts of how best to explain the data observed by Pellizzoni and colleagues (2010) since their original study conflated the contact principle and the doctrine of double effect.
Hypothesis 3 (H3). H3 predicted that the tendency for participants to align their judgments with the contact principle would be predicted by participants’ inhibitory control. This expectation was based on the view advanced by Greene and colleagues (2001) that the contact principle is a heuristic rooted in the intuitive and affective aversion to causing direct physical harm even if it results in an overall greater good.

Hypothesis 4 (H4). As H2 suggests, the predicted pattern was that participants would conform their moral judgments to the doctrine of double effect. We expected some variability with respect to this, however. H4 predicted that age would be a significant positive predictor of participants’ tendency to conform their judgments to the doctrine of double effect. This hypothesis follows from research demonstrating the relationship between sensitivity to the intentions of others and aligning one’s judgments with the doctrine of double effect (Cushman & Young, 2011). We made the reasonable assumption that older participants would be more sensitive to the systematic variations of intentionality in the dilemmas.

Hypothesis 5 (H5). H5 predicted that participants who passed the ToM tasks would be more likely to conform their moral judgments to the doctrine of double effect. This follows from previous research documenting the
relationship between sensitivity to intentions and adherence to the doctrine of double effect (Cushman & Young, 2011).

**Hypothesis 6 (H6).** Finally, H6 proposed an interaction effect between participant age and ToM abilities in predicting the tendency to conform their judgments to the doctrine of double effect. The nature of this hypothesized interaction is such that older participants were expected to align their judgments to the doctrine of double effect, but only in those cases where they have sufficiently developed ToM abilities.
METHOD

As in the study by Pellizzoni and colleagues (2010), participants were children ages 3, 4, and 5. A total of 129 children took part in the study, but 1 was subsequently excluded for being outside the desired age range. This left a final sample of $N = 128$. ($63$ female; $M_{age} = 53.2$ months, $SD = 11.4$). The distribution of participant age appears in Figure 4 below. Participants were recruited from the local community using flyers and e-mail advertisements. Data collection took place at the University of Notre Dame Center for Children and Families. Participants’ caregivers received $20$ cash in exchange for their time, and participants were given a children’s book of their choosing upon completion.
Figure 4. Summary of participants’ ages in months. The distribution is nearly uniform, but lacks participants toward the very high end of 70 months of age.

Measures

Trolley cases. Participants were presented with three of the classic trolley cases. In particular, participants rated the moral permissibility of the Footbridge case, the Loop Track case, and the Bystander case (see complete scripts for all tasks in Appendix A). The inclusion of the Footbridge case and the Loop Track case enabled the retesting of the primary hypothesis of Pellizzoni and colleagues (2010). Specifically, it allowed us to address whether or not
children ages 3, 4, and 5 make moral judgments according to the contact principle. Since the *Footbridge* case involves the protagonist intentionally causing harm through direct physical contact, and the *Loop Track* case involves the protagonist intentionally causing harm without physical contact, this is a better contrast for testing the contact principle than the method of Pellizzoni and colleagues. Their study investigated responses on the *Footbridge* case and the *Bystander* case, which differ not only in physical contact of the protagonist and the victim, but also in the intentionality of the harm to the victim. As a consequence, it is difficult to draw substantive conclusions from their findings about the extent to which children make moral judgments that align with the contact principle since their methodology conflated contact and intentionality.

To further extend the work of Pellizzoni and colleagues, participants’ responses on the *Loop Track* case as well as the *Bystander* case were contrasted in order to test the extent to which their moral judgments align with the doctrine of double effect. These two trolley cases are quite similar, but differ in one crucial aspect. As Figure 3 illustrates, the death of victim in the *Bystander* case is merely incidental since diverting the runaway trolley to the side track is saves the five on the main track regardless of the lone bystander’s presence. The same is not true for the *Loop Track* case. In this dilemma the death of the lone bystander on the side track is the essential means by which the five men on the
main track are saved. As a result of this subtle—but important—difference, it was expected that participants who aligned their judgments with the doctrine of double effect would endorse intervention in the Bystander case, but not in the Loop Track case. This would indicate a willingness to allow harm that brings about a greater good, just so long as the harm is not the means by which the greater good is achieved. This is the defining feature of the doctrine of double effect.

**Control cases.** Additionally, participants were presented with two control cases: one positive control, where intervention prevents harm without causing any additional suffering, and one negative control, where intervention produces more harm than non-intervention. The positive control cases is similar to the original Bystander case, with the exception being that there is no bystander on the side track (call this Bystander\textsubscript{absent}). Thus, in this case there are five people on the main track and no one on the side track (see Figure 5 below). Consequently, the rational decision here is to intervene and flip the switch to divert the trolley onto the side track.
Figure 5. Illustration of Bystander\textsubscript{absent} case in which the protagonist can flip the switch to divert the trolley to the empty side track. This vignette serves as a positive control, where the rational response is to intervene.

Our negative control case was also structurally similar to the original Bystander case. In keeping with the methods of Pellizzoni and colleagues, participants were presented with an inverted Bystander case (call this Bystander\textsubscript{inv}). Recall that the Bystander\textsubscript{inv} case is simply a reversal of the standard Bystander case. Instead of considering flipping a switch to save 5 people and sacrifice 1, the protagonist in the Bystander\textsubscript{inv} case is considering flipping a switch to save 1 person and sacrifice 5 (see Figure 6). This acted as our negative control vignette.
Inhibitory and effortful control. We used two measures of inhibition and effortful control: the first based on participants’ responses on a Stroop-like task, and the second based on parent report. Participants completed the happy–sad task (Laguttuta, Sayfan, & Monsour, 2011). This measures the inhibition dimension of executive functioning. Participants are presented with card (14 cm by 11 cm) depicting either a happy face or a sad face. The task requires that the participant say “happy” when presented with a sad-face card, and to say “sad” when presented with a happy-face card. There are 20 cards (10 “happy”, 10
“sad”). Correct responses are scored as 1, and incorrect responses are scored as 0. A sum score is assigned for each participant completing all 20 trials.

Additionally, participants’ caregivers completed the Effortful Control sub-scale of the Child Behavior Questionnaire-Very Short Form (CBQ-VSF; Putnam & Rothbart, 2006). This scale was designed for use with 3- to 8-year-olds. The 12-item Effortful Control sub-scale has been compared to Conscientiousness/Constraint with items derived from the original Child Behavior Questionnaire’s Inhibitory Control, Attentional Control, Low Intensity Pleasure, and Perceptual Sensitivity sub-scales.

**Receptive vocabulary task.** We assessed participants’ verbal ability using the receptive vocabulary task from the Parent Administered Language test (PAL; Stromswold, Sheffield, Truit, & Molnar, 2006). This is a picture-pointing task that presents participants with a board with 12 images of common items (e.g., mittens, helicopter, nurse). The participant must identify the item as the experimenter names it. The experimenter names 8 items, all of which are easily identifiable nouns. The script and images appear in Appendix A. The order of the items was randomized for each participant.

**Theory of mind.** Participants also completed a task designed to assess their ToM abilities. Specifically, participants’ sensitivity to the beliefs of others
was assessed using a classic contents false-belief task (Perner et al., 1987; Wellman & Liu, 2004). The complete script appears in Appendix A.

**Demographic information.** Several pieces of demographic information were collected for each participant. Specifically, his or her date of birth (for coding age in months), gender, ethnicity, parents’ levels of education, child’s bilingual ability, and household income were noted. Of these, age is the only demographic variable of particular interest to the *a priori* hypotheses, but the inclusion of additional demographic variables was done to facilitate more nuanced analyses in the case that unexpected differences emerged (e.g., gender differences).

**Procedure**

Participants arrived at the Notre Dame Center for Children and Families with their parent or legal guardian. The child was encouraged to play with available toys while the parent completed a consent form and a brief questionnaire including demographic information and the Effortful Control subscale of the CBQ-VSF. Upon completion of the questionnaire, the participant and his or her caregiver were escorted to a quiet interview room to complete the various tasks involved in the study. The room contained one small table, with two chairs. The participant’s caregiver was present and seated in a third chair
out of view of his or her child throughout the study. The experimenter was
female, and was blind to the hypotheses of the study. Sessions were videotaped
to document participants’ responses.

**Receptive vocabulary task.** Participants were introduced to the
receptive vocabulary task as the “picture game”. The participant was seated next
to the experimenter. The experimenter revealed a laminated card (28cm by 22cm)
with 12 images (see Appendix A). The experimenter then proceeded to give the
participant the instructions of the game. The participant was asked to identify
the object named by the experimenter

**False belief task.** Participants’ sensitivity to the false beliefs of others
was assessed using a contents false belief task (Perner, et al., 1987, Wellman &
Liu, 2004). Participants were shown a box of crayons. Participants were then
asked what they thought was in the box. After they responded, the experimenter
opened the box to reveal birthday candles. Next, the experimenter produced a
stuffed toy chipmunk, and introduced him as “Alvin”. The experimenter
explained that Alvin has never seen inside this crayon box, and that he had been
asleep. Next, the experimenter asked the participant what Alvin would think is
inside the crayon box (*target* question). As a check of the participant’s memory,
they were also asked if Alvin had seen inside the box (*memory* question). Passing
responses on this task were those in which the participant responded that Alvin
thinks crayons are in the box (target question), and that Alvin has not seen in the box (memory question).

**Happy-sad task.** The participant was introduced to the happy-sad task as the “opposite game”. The task began with a brief introduction to the rules, and then a practice round. The experimenter laid the cards one-at-a-time on the table in front of the participant, and the participant responded by saying “happy” when he or she saw a card displaying a sad face, or “sad” when he or she saw a card with a happy face. The practice round was identical to the testing round, with one exception; participants were corrected after an incorrect response in the practice round. The practice round continued until the participant provided four correct responses in a row. If after 20 trials the participant had not provided a sequence of four correct responses, the testing round was omitted.

The testing round proceeded just as the practice round, except that the experimenter did not immediately correct participants’ incorrect responses. Only if the participant gave four incorrect responses in a row did the experimenter remind the participant of the rules. This is in keeping with the methods of Lagattuta and colleagues (2011).

**Judgment of trolley cases.** Participants were seated next to the experimenter who presented the participant with various trolley cases outlined
above. The different vignettes were presented on a tablet computer using Prezi Viewer presentation software. Although the stimuli depicting the trolley cases were still images, this application allowed the experimenter to guide the focus to different parts of the image to coincide with the narrative. For example, when the experimenter read “Frank is on a bridge that passes over train tracks.”, the view would shift and magnify the image so as to highlight and fill the frame with the protagonist. The narrative proceeded through the “scenes” in this manner with the focus shifting to highlight the relevant subject.

For those trolley cases in which the protagonist had the option of flipping a switch to divert the trolley to a side track (i.e., the Bystander, Loop Track, Bystander\_inv, and Bystander\_absent), participants were asked, “Do you remember what the switch does? What does it do?” after the switch is first mentioned in the vignette (see Appendix A for scripts). If the participant responds incorrectly or seems confused, the experimenter reminded them the role of the switch. At the conclusion of each vignette, the experimenter asked the participant whether or not the protagonist should intervene (e.g., “What do you think Frank should do? Should Frank push the man or not push the man?”). This was the target question.

**Target question.** In order to mitigate the possibility of any recency (or primacy) effects resulting from the wording of the trolley cases’ target questions, the experimenter alternated the ordering of the pro- and anti-intervention
clauses. Thus, for a participant’s first trolley case, the experimenter would ask “Should [protagonist] [intervene] or not [intervene]?”. Then for that same participant’s second trolley case the experimenter would reverse the ordering and ask “Should [protagonist] not [intervene] or [intervene]. This alternating of the clauses in the target question continued across the three trolley cases and the two control cases.

The use of dichotomous response options on the target question has at least three advantages. First, this follows the method of Pellizzoni and colleagues (2010), and we attempted to extend their findings, staying close to the methods of the original study seemed prudent. Second, the use of a dichotomous response lowers the burden for the participants; using a Likert-type rating system would likely prove difficult for children only three years of age. Finally, related to the second point, although a Likert-type system could yield more variability in the data, and provide more avenues for potential analyses, it is uncertain that the variability would be related to the construct of interest.

The order of the trolley cases for each participant was randomized, with the constraint that the two controls always appeared at the end. The order of the two controls was also randomized. In addition, the false belief task, happy–sad task, and receptive vocabulary task were interspersed between the trolley problems such that participants did not go directly from one trolley problem to
the next. This was based on pilot testing and was intended to lessen the potential for participants’ responses on one trolley case to carry forward and influence later responses.

**Training Videos.** Prior to the presentation of each trolley case, participants were shown two brief training videos. The videos were Flash-based animations. The first depicts the behavior of trolley going down the main track; and the second depicts the behavior of the trolley when a person flips the switch to turn the trolley to the side track. Each video lasted approximately 5 seconds. Each time the training videos were presented, they were both shown a minimum of two times. Before the second presentation the experimenter asked the participant to indicate the path the trolley would take. After the participant successfully demonstrated their understanding, the experimenter moved on to the actual trolley cases.

**Task order.** A sample ordering of the various tasks is illustrated in Table 1. The rules for ordering are as follows. Every session began with the two trolley demonstration videos. These videos were always shown in the same order: main track demonstration, then side track demonstration. The training videos always preceded the presentation of a trolley case. The remaining tasks made up three groups: *trol* = trolley cases, *cog* = cognitive tasks, and *cntrl* = control trolley cases. Within each of these three groups, the orders of the specific tasks
were randomized. The tasks in both the *trol* and *cog* groups could appear in 6 different orders (i.e., 3!), while the tasks in *cntrl* could appear in 2 different orders.

<table>
<thead>
<tr>
<th>Task</th>
<th>Randomization Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Main track demonstration video</td>
<td>—</td>
</tr>
<tr>
<td>2 Side track demonstration video</td>
<td>—</td>
</tr>
<tr>
<td>3 <em>Footbridge</em> case</td>
<td><em>trol</em></td>
</tr>
<tr>
<td>4 Happy-sad task</td>
<td><em>cog</em></td>
</tr>
<tr>
<td>5 Main track demonstration video</td>
<td>—</td>
</tr>
<tr>
<td>6 Side track demonstration video</td>
<td>—</td>
</tr>
<tr>
<td>7 <em>Bystander</em> case</td>
<td><em>trol</em></td>
</tr>
<tr>
<td>8 Receptive vocabulary task</td>
<td><em>cog</em></td>
</tr>
<tr>
<td>9 Main track demonstration video</td>
<td>—</td>
</tr>
<tr>
<td>10 Side track demonstration video</td>
<td>—</td>
</tr>
<tr>
<td>11 <em>Loop Track</em> case</td>
<td><em>trol</em></td>
</tr>
<tr>
<td>12 False-belief task</td>
<td><em>cog</em></td>
</tr>
<tr>
<td>13 Main track demonstration video</td>
<td>—</td>
</tr>
<tr>
<td>14 Side track demonstration video</td>
<td>—</td>
</tr>
<tr>
<td>15 <em>Bystander</em> <em>inv</em> case</td>
<td><em>cntrl</em></td>
</tr>
<tr>
<td>16 <em>Bystander</em> <em>absent</em> case</td>
<td><em>cntrl</em></td>
</tr>
</tbody>
</table>

The full sequence of tasks took approximately 15 minutes to complete.

After which the participants were invited to select a children’s book to take home with them as a gift. The participant’s caregiver was also given $20 in cash and debriefed.
Plan of Analysis

In the current project we adopted a Bayesian approach to model fitting; this has several advantages over the classical “frequentist” methods. Perhaps the strongest argument for a Bayesian approach is the fact that Bayesian inference allows for the incorporation of prior information in the estimation of parameters. And depending on the nature of this evidence, this can be done in a very flexible manner. Additionally, the Bayesian approach handles missing data in a very natural way as part of the estimation process (Gill, 2009). Bayesian inference also allows for more intuitive interpretation of confidence intervals associated with parameter estimates. This is in sharp contrast to the classical approach, which relies on long-run probabilities, and therefore, confidence limits based on repeated sampling. Lastly, the Bayesian approach is very well suited to multilevel modeling (Gelman & Hill, 2007).
RESULTS

Data analysis was completed in the R statistical programming language (R Core Team, 2013). All model fitting was done using the Gibbs sampler JAGS (Plummer, 2003) and the rjags package (Plummer, 2014). The complete R code and JAGS model code both appear in Appendix B.

Preliminary Data Analysis

Exclusionary criteria. Of the total sample of 128, a considerable number failed at least one of the control trolley cases (i.e., Bystander_{inv}, or Bystander_{absent}). These participants were excluded from our primary analyses. Thus, our primary research questions are addressed using only the subsample of 79 (41 female, \( M_{age} = 56.19, SD = 11.94 \)) participants not failing a control trolley case.

Missing data mechanism. The remaining data constitutes an \( N \times p \) matrix, where \( N \) is our remaining sample of 79, and \( p \) is our number of variables.

\(^2\) Participant age was a significant negative predictor of the tendency to fail at least one of the control cases, \( b = -0.08 \) (0.02), 95\% HPD interval = [–0.11, –0.04].
10. In this complete matrix, there was a missingness rate of only 4.3%. Figure 7 illustrates the pattern of missingness in the $N \times p$ matrix. Figure 7 is sorted by age, with younger participants’ records appearing lower in the figure. Red blocks represent missing observations, while black, white, and grayscale blocks represent observed data—with higher values on a variable denoted by darker shades. Age was used as the sorting variable because of our $a priori$ interest in the role of age as a predictor, and also because we observed during data collection that the younger participants tended to struggle with the happy-sad task. This can be seen in Figure 7 where we note more missingness on the happy-sad task in the lower portion of that column (i.e., for younger participants). We also note more missingness on the outcome variables for younger participants (i.e., contact, and dde). This is again illustrated by the greater number of red cells in the lower half of those columns. Thus, we are inclined to believe that the missingness mechanism is missing at random (MAR). That is, the missingness on the outcome variable is related not to the outcome variable itself, but rather to a measured predictor variable (i.e., age_m). The regression models discussed below all use age_m and vocab as auxiliary variables to model the missing data.
Figure 7. Summary of missing data in $N \times p$ matrix of variable sorted by age in months. Red blocks indicate missing observations. Black, white, and grayscale blocks represent observed data, with higher values on a variable are denoted by darker shades. This figure is sorted by the variable age_m, such that younger participants’ records appear lower on figure.

Testing A Priori Hypotheses

Conforming to contact principle. Recall that H1 was concerned with replicating the findings of Pellizzoni and colleagues—only with modified stimuli. That is, we anticipated that participants would conform their judgments to the
contact principle, which suggests that causing harm through physical contact is considered morally worse than causing equivalent harm without physical contact. Thus, H1 asserted that participants would be particularly likely to endorse intervention in the *Loop Track* case and not in the *Footbridge* case.

With respect to participants’ pattern of judgments, any single participant’s judgments can take one of 4 patterns: (1) endorse intervention in the *Loop Track* case but not in the *Footbridge* case, (2) endorse intervention in the *Footbridge* case but not in the *Loop Track* case, (3) endorse intervention in both cases, and (4) endorse intervention in neither case. The first of these is the one indicative of conforming to the contact principle; therefore, this was the pattern that of interest to us. We treated participants’ tendency to conform or not conform to the contact principle as a series of independent Bernoulli trials where Pattern 1 represents a “success” and Patterns 2 through 4 represent a “failure”.

Participants who showed response Pattern 1 received a score of “1” for conforming to the contact principle. Those with Patterns 2 through 4 received as score of “0”. We were interested in $\theta_{contact}$, the probability of conforming one’s pattern of judgments to the contact principle. We used a flat prior of $Beta(1, 1)$ to estimate the posterior distribution of $\theta_{contact}$. We ran 3 Markov chains for
10,000 draws to estimate the posterior distribution following a burn-in period of 1000.

Figure 8 illustrates the distribution of the posterior estimate of $\theta_{\text{contact}}$, the mean of the distribution, as well as the highest posterior density interval\(^3\) (HPD), and the probability that $\theta_{\text{contact}}$ falls above or below 0.25. Since there were 4 potential patterns of responses to the Footbridge case and the Loop Track case, 0.25 represents responding at chance. As this figure suggests, our estimate of $\theta_{\text{contact}}$ is 0.22 (0.05), 95% HPD interval = [0.13, 0.32], with a 75.4% probability of $\theta_{\text{contact}}$ being greater than 0.25. Thus, H1 is not supported by these data according to the traditional conventions of statistical significance.

\(^3\) HPD intervals in the Bayesian framework are similar to standard confidence intervals in the classical frequentist approach. For hypothesis testing, we reject $H_0$ when the value it predicts is not contained within the interval.
Figure 8. Posterior distribution of $\theta_{contact}$ suggesting that the probability of conforming to the contact principle is not different than what would be expected with participants responding at chance.

Conforming to doctrine of double effect. Our second a priori hypothesis, H2, asserted that participants would conform their moral judgments on the trolley cases to the pattern prescribed by the doctrine of double effect. This suggests that harm that brings about a greater good may be permissible provided that the harm is not the necessary means by which that greater good is achieved. H2 was tested by contrasting participants’ judgments on two trolley cases: the Loop Track case, and the Bystander case. According to the doctrine of
double effect, the intervention in the *Bystander* case is considered permissible, while intervention in the *Loop Track* case is considered forbidden.

Just as in H1, participants’ responses could only haven take one of 4 patterns: (1) endorse intervention in the *Bystander* case but not in the *Loop Track* case, (2) endorse intervention in the *Loop Track* case but not in the *Bystander* case, (3) endorse intervention in both cases, and (4) endorse intervention in neither case. The first of these patterns is demonstrative of adherence to the doctrine of double effect. We will consider this pattern a success, and patterns (2) through (4) a failure. And again, we treat participants’ patterns of responses as a set of independent Bernoulli trials with equal probability.

Those participants who displayed response Pattern 1 received a score of 1 for conforming to the contact, while those who displayed Patterns 2 through 4 received as score of 0. We were interested in $\theta_{dde}$, the probability of conforming one’s judgments to the doctrine of double effect. We estimated the posterior distribution of $\theta_{dde}$ using the flat prior $Beta(1, 1)$. We ran 3 chains for 10,000 draws from the posterior distribution after a burn-in period of 1000.

Figure 9 illustrates the distribution of the posterior estimate of $\theta_{dde}$, the mean of the distribution, the HPD interval, and the probability that $\theta_{dde}$ falls above or below 0.25, which represents chance. As this figure suggests, our
estimate of $\theta_{dde}$ is 0.12, 95% HPD interval = $[0.05, 0.20]$, and suggests there is just a 99.7% probability of $\theta_{dde}$ being less than 0.25.

![Figure 9](image.png)

*Figure 9.* Posterior distribution of $\theta_{dde}$ suggesting that the probability of conforming to the doctrine of double effect is less than what would be expected with participants responding at chance.

**Decision to level.** The sampling process in this study did not exclude the possibility of multiple siblings from one family taking part. Of the 79 participants in the sample, 25 were part of a sibling pair or triad. As a result, we might be concerned about violating the assumption of the independence of errors.
in regression. Thus, prior to moving on to fitting regression models, we first checked the intra-class correlation coefficient (ICC) of fitted vary-intercept multilevel logistic regression models with participants nested in families. Equation 1 below illustrates the computation of the ICC for multilevel logistic regression models (Kreft & de Leeuw, 1998).

\[
ICC = \left( \frac{\sigma_u^2}{\sigma_u^2 + \frac{\pi^2}{3}} \right) 
\]  

Where \( \sigma_u^2 \) is the estimated error variance of the level-2 equation for intercept, and \( \pi \) is the mathematical constant 3.14. The value of the ICC is our estimate of the extent to which between-group variance dominates over within-group variance. For our purposes, a high ICC value would suggest we consider conducting our analyses using a multilevel approach to account for the similarity between siblings.

Recall our two primary outcome variables of interest are whether the participant conformed their judgments to (1) the contact principle and (2) the doctrine of double effect. Therefore, the ICC was calculated using varying-intercept multilevel logistic regression models for both of these dependent variables.
Regression models were fitted with $N(0, 33.33)$ priors on the level-2 regression parameters, and standard half-$t$ priors with $\nu = 5$ on the level-2 variance parameter (Gelman, 2006; Polson & Scott, 2011). We ran 3 Markov chains of length 50,000 and used an initial burn-in period of 10,000 and a thinning interval of 5. Since posterior distributions of estimates of variance are bounded at 0 and skewed, we used the mode of the posterior distributions, rather than the mean to calculate the ICC. In both cases, we found the ICC not to be overly large. In the case of conforming to the contact principle we found an ICC = .17, 95% HPD interval = [.00, .46], and in the case of conforming to the doctrine of double effect, ICC = .10, 95% HPD interval = [.00, .39]. These ICC values were sufficiently high—and the credible intervals sufficiently wide—that we elected to use multilevel modeling, with participants nested within families (Kreft & de Leeuw, 1998). Although we were not interested in family-level predictors, this approach accounted for the non-independence of observations.

Inhibitory control and the contact principle. Our third hypothesis, H3, concerned the possible predictors of the tendency of participants to conform their judgments to the contact principle. Specifically, we tested the hypothesis that inhibition and effortful control would be negative predictors of the tendency to conform to the contact principle. This was tested using binomial logistic regression with participants’ status (i.e., did not conform = 0, conformed = 1)
regressed on their scores on the happy-sad task, as well as parent-report effortful control scores. Note that we intentionally collapsed the three patterns of judgment that are not indicative of conforming to the doctrine of double effect, and thus, opted to use binomial rather than multinomial logistic regression. This is done for two reasons. First, it facilitates the interpretation of the resulting parameter estimates. And just as importantly, this also represents a more direct test of our a priori hypothesis. Mean-centered age in months was also included as a predictor. We also used age and vocabulary ability as auxiliary variables to predict missingness in the predictors of the primary regression model. Equation 2 gives the multilevel model used to test H3.

$$\logit(\pi_{ij}) = \beta_{0j} + \beta_1 \times \text{age}_i + \beta_2 \times \text{inhib\_cntrl}_i + \text{eff\_cntrl}_i,$$

$$\beta_{0j} = \gamma_{00} + \zeta_j$$ (2)

where $i$ indexes individual participants and $j$ indexes family. And where $\beta$ and $\gamma$ are regression coefficients, while $\zeta$ denotes the level-2 error term.

The model was fitted using $N(0, 33.33)$ priors on the regression parameters and a half-$t$ prior with $\nu = 5$ on the level-2 variance component. We used three Markov chains of length 50,000 to fit the model, with a thinning interval of 5, and a burn-in period of 1000. The results of this model appear in
Table 2. As expected, scores on the happy-sad task, which indexes inhibitory control, were negatively predictive of participants’ tendencies to conform to the contact principle, $b = -0.23 \ (0.11)$, 95% HPD interval $= [-0.46, -0.02]$. Neither age, nor effortful control proved to be significant predictors of the tendency to conform one’s judgments to the contact principle.

**TABLE 2**

**INHIBITORY CONTROL AND EFFORTFUL CONTROL PREDICTING CONFORMING TO CONTACT PRINCIPLE**

<table>
<thead>
<tr>
<th>M1a</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mode</td>
<td>HPD\text{lower}</td>
<td>HPD\text{upper}</td>
</tr>
<tr>
<td>(intercept)</td>
<td>$\gamma_{00}$</td>
<td>$-2.40$</td>
<td>$0.59$</td>
<td>$-2.33$</td>
<td>$-3.61$</td>
</tr>
<tr>
<td>Age</td>
<td>$\beta_1$</td>
<td>$0.06$</td>
<td>$0.05$</td>
<td>$0.05$</td>
<td>$-0.04$</td>
</tr>
<tr>
<td>Inhib. cntrl</td>
<td>$\beta_2$</td>
<td>$-0.23$</td>
<td>$0.11$</td>
<td>$-0.22$</td>
<td>$-0.46$</td>
</tr>
<tr>
<td>Eff. cntrl</td>
<td>$\beta_3$</td>
<td>$0.02$</td>
<td>$0.07$</td>
<td>$0.02$</td>
<td>$-0.12$</td>
</tr>
<tr>
<td>$\sigma_0^2$</td>
<td></td>
<td>$1.06$</td>
<td>$0.87$</td>
<td>$0.63$</td>
<td>$0.00$</td>
</tr>
</tbody>
</table>

$M_{\text{deviance}}$ | 2014.00 |
PSRF | 1.01 |

**M1a convergence diagnostics.** The model’s convergence was assessed in two ways. First, we used graphical checks of the trace plots from the Markov chain Monte Carlo (MCMC) draws from the posterior distribution. Figure 10 illustrates the trace plots from the above model. These all show good mixing of
MCMC chains, and are strongly suggestive of model convergence. We also calculated the potential scale reduction factor (PSRF) suggested by Gelman and Rubin’s (1992). A PSRF close to 1 is indicative of model convergence, and values over 1.10 are indicative of models that have not converged.

Figure 10. Summary of trace plots from model M1a above. These illustrate fast mixing of chains, and overall model convergence.
**Sensitivity analysis of M1a.** In order to assess the influence of the priors, we next proceeded to conduct sensitivity analysis of this model by comparing the results obtained in M1a to the results obtained using different priors on the regression parameters. We also varied the link function to test the model using different versions of the so-called robit link, which is robust to outlying observations (Liu, 2006). Thus, we re-ran the same model using 3 different combinations of priors and link functions. In M1b we used \(N(0, 1.0E+4)\), priors on the regression parameters, the logit link function, and a half-\(t\) with \(\nu = 5\) on the level-2 variance component. M1c used \(N(0, 1.0E+4)\) priors on the regression parameters as well as the standard Cauchy as a link function—a robit link—and a half-\(t\) with \(\nu = 5\) as the prior for the level-2 variance component. Finally, M1d used \(t\)-distribution priors with \(\nu = 3\) on the regression parameters, a half-Cauchy prior on the level-2 variance component, and a \(t\)-distribution (robit) link function with \(\nu = 3\). The results of these models appear in Table 3, and the combination of link functions and priors appear in Table 4. These combinations will be used again in the sensitivity analyses of models testing H4 through H6.
TABLE 3
MODELS PREDICTING TENDENCY TO CONFORM TO CONTACT PRINCIPLE

<table>
<thead>
<tr>
<th></th>
<th>( \gamma_{00} ) Mean (SD)</th>
<th>( \beta_1 ) Mean (SD)</th>
<th>( \beta_2 ) Mean (SD)</th>
<th>( \beta_3 ) Mean (SD)</th>
<th>( \sigma_0^2 ) Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1a</td>
<td>-2.40 (0.59)</td>
<td>0.06 (0.05)</td>
<td>-0.23 (0.11)</td>
<td>0.02 (0.07)</td>
<td>0.63</td>
</tr>
<tr>
<td>M1b</td>
<td>-2.39 (0.59)</td>
<td>0.06 (0.05)</td>
<td>-0.22 (0.11)</td>
<td>0.01 (0.07)</td>
<td>0.66</td>
</tr>
<tr>
<td>M1c</td>
<td>-5.84 (3.23)</td>
<td>0.09 (0.13)</td>
<td>-0.67 (0.48)</td>
<td>0.06 (0.18)</td>
<td>0.57</td>
</tr>
<tr>
<td>M1d</td>
<td>-1.77 (0.55)</td>
<td>0.04 (0.04)</td>
<td>-0.16 (0.10)</td>
<td>0.01 (0.06)</td>
<td>0.48</td>
</tr>
</tbody>
</table>

TABLE 4
PRIORS AND LINK FUNCTIONS USED IN SENSITIVITY ANALYSES

<table>
<thead>
<tr>
<th></th>
<th>Priors</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Logit</td>
</tr>
<tr>
<td>b</td>
<td>Logit</td>
</tr>
<tr>
<td>c</td>
<td>Standard Cauchy</td>
</tr>
<tr>
<td>d</td>
<td>t(( v = 3 ))</td>
</tr>
</tbody>
</table>

As Table 3 suggests, the sensitivity analysis reveals some variability due to the choice of model priors and link function. In particular, when using the standard Cauchy link function in M1c, and when using \( t \)-distribution priors on regression parameters, half-Cauchy on the level-2 variance parameter, and a \( t \)-distribution link function in M1d we see parameter estimates that are quite a bit
different than M1a. Moreover, in both M1c and M1d we observed that the effect of inhibitory control was no longer a significant predictor of the tendency to conform one’s judgments to the contact principle. Note that in both M1c and M1d we used a robit link function. The differences between these and the logit models might indicate that outliers are influencing the observed effect of inhibitory control on participants’ tendency to conform to the contact principle. Furthermore, the sensitivity analysis suggests we ought to be at least a bit cautious in the interpretation of the original results of M1a.

**Post-hoc test.** A follow-up test suggested that the effect of inhibitory control on the tendency not to conform to the contact principle might have been at least somewhat driven by a positive relationship between inhibitory control and endorsing intervention in the *Footbridge* case, $b = 0.18$ (0.11); 95% HPD interval = $[-0.02, 0.42]$. And although this relationship does not quite reach traditional significance threshold, the HPD interval, and the posterior distribution illustrated in Figure 11 seems to suggest there is likely to be some effect in the positive direction.
Figure 11. Posterior distribution of the effect of inhibitory control on the tendency to endorse intervention in the footbridge case.

**Age and the doctrine of double effect.** Recall that according to H4 participant age would be a significant positive predictor of the tendency to make moral judgments that align with the doctrine of double effect. This hypothesis was tested using multilevel logistic regression. And again, the dependent variable in model was participants’ patterns of judgments, which was coded dichotomously (i.e., did not conform = 0, conformed = 1).

The only predictor in this model was participant age in months (mean-centered). Equation 3 describes the multi-level model. The model was fitted using
\(N(0, 33.33)\) priors on the regression parameters and a half-\(t\) prior with \(\nu = 5\) on the level-2 variance component. We used three Markov chains of length 50,000 to fit the model, with a thinning interval of 5, and a burn-in period of 1000. The results of this model appear in Table 5. These results show a significant negative effect of age on participants’ tendency to align their judgments with the doctrine of double effect, \(b = -0.11\ (0.05), 95\%\ HPD\ interval = [-0.21, -0.02]\. This is in the opposite direction as was predicted in H4.

\[
\logit(\pi_y) = \beta_{0j} + \beta_1 \times \text{age}_i \\
\beta_{0j} = \gamma_{00} + \zeta_j
\]  

(3)

<table>
<thead>
<tr>
<th>TABLE 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE PREDICTING CONFORMING TO DOCTRINE OF DOUBLE EFFECT</td>
</tr>
<tr>
<td>M2a</td>
</tr>
<tr>
<td>(intercept)</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>(\sigma_{0}^2)</td>
</tr>
<tr>
<td>(M_{\text{deviance}})</td>
</tr>
<tr>
<td>PSRF</td>
</tr>
</tbody>
</table>

59
**M2a convergence diagnostics.** Model convergence was again assessed through graphical checks of the trace plots from the Markov chains (Figure 12), and using Gelman and Rubin’s PSRF (1992). Both the trace plots and the PSRF indicate the model converged appropriately.

*Figure 12.* Trace plots for parameter estimates in M2a. These suggest fast mixing of Markov chains, and that the model converged appropriately.
Sensitivity analysis of M2a. The influence of the priors was assessed through sensitivity analysis. We used the combinations of priors and link functions described in Table 4. Once again, the parameter estimates derived from the different models seem in general agreement, but we note that there is some variability in the intercept parameter—particularly for the robit model M2c, which used the standard Cauchy as link function. It is also noteworthy that his model required twice the iterations and thinning interval to achieve convergence. And furthermore, the effect of age did not quite surpass the threshold of significance in M2c.

**TABLE 6**

MODELS WITH AGE PREDICTING DOCTRINE OF DOUBLE EFFECT

<table>
<thead>
<tr>
<th></th>
<th>( \gamma_0 ) Mean (SD)</th>
<th>( \beta_1 ) Mean (SD)</th>
<th>( \sigma_0^2 ) Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2a</td>
<td>-2.96 (0.70)</td>
<td>-0.11 (0.05)</td>
<td>0.59</td>
</tr>
<tr>
<td>M2b</td>
<td>-2.97 (0.68)</td>
<td>-0.11 (0.05)</td>
<td>0.62</td>
</tr>
<tr>
<td>M2c</td>
<td>-5.51 (2.76)</td>
<td>-0.20 (0.15)</td>
<td>0.54</td>
</tr>
<tr>
<td>M2d</td>
<td>-2.18 (0.64)</td>
<td>-0.08 (0.04)</td>
<td>0.43</td>
</tr>
</tbody>
</table>

ToM and the doctrine of double effect. According to H5, we expected to find a significant positive relationship between the tendency to pass
the two ToM tasks and the tendency to conform to the doctrine of double effect.

This hypothesis was tested using binomial logistic regression. Our model predicted the tendency to conform to the doctrine of double effect; this was coded dichotomously (i.e., did not conform = 0, conformed = 1).

The primary predictor in this model was a dichotomous variable indicating whether or not participants had passed the ToM task. The multilevel model is described in Equation 4 below. Age and vocabulary ability were used as auxiliary variables to model missing data in the predictor. The model was fitted using $N(0, 33.33)$ priors on the regression parameters and a half-$t$ prior with $\nu = 5$ on the level-2 variance component. We used three Markov chains of length 50,000 to fit the model, with a thinning interval of 5, and a burn-in period of 1000. The results of this model appear in Table 7. As these results indicate, ToM was not a significant predictor of the tendency of participants to conform their judgments to the doctrine of double effect.

$$\logit \left( \pi_j \right) = \beta_{0j} + \beta_1 \times \text{ToM}_i$$

$$\beta_{0j} = \gamma_{00} + \zeta_j$$
TABLE 7

ToM PREDICTING CONFORMING TO DOCTRINE OF DOUBLE EFFECT

<table>
<thead>
<tr>
<th>M3a</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
<th>HPDlower</th>
<th>HPDupper</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td>γ00</td>
<td>-2.07</td>
<td>0.55</td>
<td>-2.01</td>
<td>-3.15</td>
</tr>
<tr>
<td>ToM</td>
<td>β1</td>
<td>-1.21</td>
<td>1.01</td>
<td>-1.10</td>
<td>-3.23</td>
</tr>
<tr>
<td></td>
<td>σ_θ²</td>
<td>0.89</td>
<td>0.8</td>
<td>0.45</td>
<td>0.00</td>
</tr>
</tbody>
</table>

M3a convergence diagnostics. The convergence of model M3a was assessed using Gelman and Rubin’s PSRF (1992), as well as graphical check of the mixing of Markov chains. Figure 13 below illustrates trace plots of the Markov chains for this model, and is suggestive of fast mixing of chains and appropriate model convergence.
Figure 13. Trace plots of Markov chains for model M3a. These illustrate fast mixing and appropriate convergence.

**Sensitivity analysis of M3.** In order to investigate the robustness of the results of M3a, we conducted a sensitivity analysis in which we varied the link function as well as the priors on regression coefficients. The combination of priors and link functions used are described in Table 4 above. The results appear in Table 8 below. We see good consistency of parameter estimates between
models, with the exception M3c, which used the standard Cauchy as the link function. This model shows the most deviation from other models, but in none of these models was the significance of a parameter estimate impacted by the choice of regression priors or link function.

TABLE 8
MODELS FOR ToM PREDICTING DOCTRINE OF DOUBLE EFFECT

<table>
<thead>
<tr>
<th>Model</th>
<th>$\gamma_{00}$ (Mean, SD)</th>
<th>$\beta_1$ (Mean, SD)</th>
<th>$\sigma_0^2$ (Mode)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M3a</td>
<td>-2.07 (0.55)</td>
<td>-1.21 (1.01)</td>
<td>0.45</td>
</tr>
<tr>
<td>M3b</td>
<td>-2.07 (0.55)</td>
<td>-1.23 (1.01)</td>
<td>0.48</td>
</tr>
<tr>
<td>M3c</td>
<td>-3.27 (1.70)</td>
<td>-10.31 (11.41)</td>
<td>0.48</td>
</tr>
<tr>
<td>M3d</td>
<td>-1.54 (0.48)</td>
<td>-0.82 (0.68)</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Age, ToM, and the doctrine of double effect. Recall that the final hypothesis concerned the interactive effect of participant age and ToM abilities. In particular, H6 asserted that participant age and ToM abilities would interact to predict participants’ tendencies to conform their judgments to the doctrine of double effect. We expected the nature of this interaction would be such that for those individuals passing the ToM task, age would not be an important predictor. This was again tested using binomial logistic regression, with the model predicting the tendency to conform to the doctrine of double effect (i.e.,
did not conform = 0, conformed = 1). Equation 5 describes the multilevel model.

\[
\text{logit}(\pi_{ij}) = \beta_{0j} + \beta_1 \times \text{age}_i + \beta_2 \times \text{ToM}_i + \beta_3 \times (\text{age}_i \times \text{ToM}_i) \\
\beta_{0j} = \gamma_{00} + \zeta_j
\]

(5)

We used \( N(0, 33.33) \) priors on the regression parameters and a half-\( t \) prior with \( \nu = 5 \) on the level-2 variance component. Markov chains of length 50,000 were used to fit the model, with a thinning interval of 5, and a burn-in period of 1000. The results of this model appear in Table 9. As this table illustrates, the interaction of age and ToM was not a significant predictor of conforming to the doctrine of double effect, \( b = -0.06 (0.12), \) 95\% HPD interval = \([-0.19, 0.31]\).

However, we again find that age is a significant negative predictor of conforming to the doctrine of double effect, \( b = -0.18 (0.09), \) 95\% HPD interval = \([-0.36, -0.01]\).
### TABLE 9

**AGE × ToM PREDICTING CONFORMING TO DOCTRINE OF DOUBLE EFFECT**

<table>
<thead>
<tr>
<th>M4a</th>
<th>Mean</th>
<th>SD</th>
<th>Mode</th>
<th>HPD_{lower}</th>
<th>HPD_{upper}</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td>$\gamma_0$</td>
<td>-3.61</td>
<td>1.18</td>
<td>-3.33</td>
<td>-5.97</td>
</tr>
<tr>
<td>Age</td>
<td>$\beta_1$</td>
<td>-0.18</td>
<td>0.09</td>
<td>-0.16</td>
<td>-0.36</td>
</tr>
<tr>
<td>ToM</td>
<td>$\beta_2$</td>
<td>0.59</td>
<td>1.52</td>
<td>0.52</td>
<td>-2.39</td>
</tr>
<tr>
<td>Age × ToM</td>
<td>$\beta_3$</td>
<td>0.06</td>
<td>0.12</td>
<td>0.05</td>
<td>-0.19</td>
</tr>
<tr>
<td>$\sigma_0^2$</td>
<td></td>
<td>1.31</td>
<td>1.24</td>
<td>0.67</td>
<td>0.00</td>
</tr>
</tbody>
</table>

$M_{\text{deviance}}$ 2403.00

\[
\text{PSRF} \quad 1.02
\]

**M4a convergence diagnostics.** The convergence of model M4a was assessed by through graphical checks of the Markov chains, as well as Gelman and Rubin’s (1992) PSRF. The trace plots for parameter estimated in M4a appear in Figure 14 below. These illustrate fast mixing of Markov chains, and good convergence of the model.
Figure 14. Summary of trace plot for Markov chains used to estimate parameters in M4a. These illustrate the model converged appropriately.

**Sensitivity analysis of M4a.** We conducted a sensitivity analysis to investigate the effect of the choice of priors and link function on the results of model M4a. We fitted 3 additional regression models using the combination of priors and link functions illustrated in Table 4. The results of these models, and the original model, M4a, appear in Table 10.
Of these models, we see that parameter estimates of M4c do appear to deviate from the other models, but we also see variability in M4d. Recall that M4c used the Cauchy link function recommended by Liu (2006) as a robust alternative to the logit link, and M4d used the half-Cauchy prior on the level-2 variance parameter. Despite some variability, in none of these alternate models was the significance of parameter estimates impacted. These results suggest that the parameter estimates obtained in M4a are generally consistent with other reasonable priors and regardless of logit or robit models, but that since the sample size is still small, the choice of priors still impacts the posterior estimates noticeably.

### TABLE 10

MODELS FOR AGE × ToM PREDICTING DOCTRINE OF DOUBLE EFFECT

<table>
<thead>
<tr>
<th></th>
<th>$\gamma_{00}$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mode</td>
</tr>
<tr>
<td>M4a</td>
<td>-3.61 (1.18)</td>
<td>-0.18 (0.09)</td>
<td>0.59 (1.52)</td>
<td>0.06 (0.12)</td>
<td>0.67</td>
</tr>
<tr>
<td>M4b</td>
<td>-3.52 (1.13)</td>
<td>-0.17 (0.09)</td>
<td>0.48 (1.48)</td>
<td>0.05 (0.12)</td>
<td>0.70</td>
</tr>
<tr>
<td>M4c</td>
<td>-6.72 (4.59)</td>
<td>-0.26 (0.48)</td>
<td>-16.24 (16.77)</td>
<td>0.08 (1.02)</td>
<td>0.57</td>
</tr>
<tr>
<td>M4d</td>
<td>-2.37 (0.92)</td>
<td>-0.09 (0.06)</td>
<td>-0.32 (0.82)</td>
<td>0.00 (0.09)</td>
<td>0.32</td>
</tr>
</tbody>
</table>
DISCUSSION

The present study makes several important contributions. Most obviously, we build on the work investigating the extent to which individuals align their judgments with the contact principle the doctrine of double effect (e.g., Greene et al., 2001, Mikhail, 2000, Pellizzoni et al., 2010, Stey et al., 2013).

However, despite its merits, the present study did not replicate the findings of Pellizzoni and colleagues (2010); we found little evidence of a strong tendency for children between the ages of 3 and 5 to align their moral judgments with the contact principle. Furthermore, there was also no strong evidence indicating a tendency for children in this age group to conform their judgments to the doctrine of double effect, which would have been a completely novel finding. In fact, these results seem to suggest that children’s tendency to conform—or not conform—to these heuristic principles is no greater than we would expect by chance. And in the case of the doctrine of double effect, we see participants demonstrating this pattern significantly less than we would expect by chance. What should we make of this result?
There are a variety of possible reasons that this project and the work of Pellizzoni and colleagues (2010) reached different conclusions. The most apparent point of divergence between the two studies were the methods used to present the trolley cases as well as the wording of the target question for each vignette. The current project presented all the trolley cases to the participant in one session while Pellizzoni and colleagues presented the Bystander case and the Footbridge case in separate sessions 1 month apart. Additionally, the current project used a tablet computer to present the training videos and the test stimuli, while the studies reported by Pellizzoni and colleagues used Lego™ models in combination with photographs of two candidate outcomes as the test stimuli. Still, it is not immediately obvious why these two different methods would produce differing results.

A potential explanation is that the Lego™ models used by Pellizzoni and colleagues were more engaging, and thus, participants might have been more inclined to think deeply about the dilemmas. Of course, this is merely speculative. Furthermore, the literature on dual representations (e.g., DeLoache, 2000) and perceptual salience of symbols (Petersen & McNeil, 2013) might even suggest that the opposite is true—that use of Lego™ models might have had a deleterious impact on participants’ understanding of the models as symbolic.
Consequences for Competing Theories

The present study can be understood as pitting two competing theories against one another. According to Greene’s view, moral judgments are the product of a dual process model in which characteristically utilitarian judgments (e.g., flipping the switch in the Bystander case) are produced through deliberate reasoning, while deontological judgments arise from affective aversions to inflicting direct physical harm (e.g., not pushing the man in the Footbridge case). Mikhail’s view, alternatively, suggests that our moral judgments are a function of our sensitivity to the intentions, beliefs, and desires of others. Mikhail’s likens the our moral sense to that of an intuitive lawyer—making judgments of blameworthiness and punishment based on things like counts of battery and the intentionality involved in causing harm.

Status of Greene’s account. For Greene’s view to be vindicated, we would have expected that participants align their judgments with the contact principle. That is, we would have expected them to readily endorse intervention in the Loop Track case, but not in the Footbridge case. Once again, we found no evidence suggesting that participants as a whole were displaying this pattern more frequently than would be expected by chance.

We did find some evidence supporting Greene’s view, however, in our test of H3. Recall that we had predicted participants with greater inhibitory control
would be less likely to conform their judgments to the contact principle. Our reasoning was based Greene’s claim that the contact principle is guided by an affective aversion to causing direct physical harm. Consequently, it is not surprising that participants with greater inhibitory control were also less likely to align their judgments with the contact principle. In fact, this follows quite naturally, but it is an interesting result, nonetheless. Furthermore, our follow-up tests seemed to suggest that this effect was driven by a greater willingness to endorse intervention in the Footbridge case for those participants with greater inhibitory control. This again aligns with what Greene’s view might have predicted. Participants who were better able to inhibit their prepotent responses on the happy-sad task were more inclined to do what others might find morally impermissible—sacrificing the victim through direct physical contact. Thus, while there does not seem to be a strong bias for children ages 3, 4, and 5 to align their judgments with the contact principle, it seems those who do so also tend to have less inhibitory control.

**Status of Mikhail’s account.** For Mikhail’s view to be supported, we would have expected to find that participants were aligning their judgments with the doctrine of double effect. Meaning, we would have expected that participants display a pattern of endorsing intervention in the Bystander case, but not endorsing intervention in the Loop Track case. And again, we found no
evidence suggesting that participants as a whole aligned their judgments in this way more often than what would be expected by chance. In fact, we found an effect in the opposite direction—participants displaying this pattern at levels less than would be expected by chance. It is difficult to know what to make of this finding, other than to note that children of this age group do not yet show this bias. Furthermore, it was also illuminating to see that older participants were even less likely to show the pattern of judgments characteristic of the doctrine of double effect. This is even more of a surprising result, since we would have expected older children to behave more like adults—who do tend to display this bias. Here again, we can only speculate about the nature of the this effect. But one possibility is that the tendency to conform one’s judgments to the doctrine of double effect perhaps develops along a u-shaped function.

Implications and Future Directions

The inability to reproduce the results obtained by Pellizzoni and colleagues (2010) is both vexing and interesting. One important avenue for future work is to untangle the specific mechanisms that might have contributed to the asymmetry between the present study and the work of Pellizzoni and colleagues. One candidate explanation—the salience of symbols—was noted above, but it is also important to once again mention the differences between
how responses to the vignettes were solicited in the two studies. The current study asked explicitly for participants to indicate which option the protagonist ought to pursue (e.g., push the man, or not push the man). The studies reported by Pellizzoni and colleagues asked participants to indicate their responses by pointing to one of two photos illustrating the two possible outcomes (i.e., 5 people saved, or 5 people not saved). Perhaps by lessening the burden of task demands, Pellizzoni and colleagues were able to detect more subtle effects. But we must again consider whether these are meaningful, since their approach did not require participants to explicitly indicate whether intervention or non-intervention was preferred. Instead, they merely indicated the preferred outcome, and then the researchers assumed that the selection was based on a preference for a given action.

**Broader issues of measurement.** Like the majority of the work in the social sciences, research on moral cognition is condemned to assess constructs of interest through indirect measurements. By their very nature, these indirect measurements are imperfect. The present study employed the famous trolley cases because they have become a standard methodological tool in the field. But it is important that we question whether or not these stimuli are appropriate for this age group. The findings reported in this paper are largely incongruent with the work reported by Pellizzoni and colleagues, and we must wonder which result
represents the closer approximation to the truth. Given the slight differences in methods of measurement, we are inclined to conclude that whatever the true effect may be, it is rather subtle and also sensitive task demands.

**Developmental theory.** We began this paper by noting that the recent revival of work in moral psychology has been driven by contributions from increasingly diverse fields. It is also important to note that the field of moral psychology arose largely because of work by developmentalists (e.g., Piaget 1965; Kohlberg, 1969), and we are still inclined to think of moral judgment and decision making as a developmental construct. Thus, although the recent contributions from diverse fields have been hugely important, it is crucial to bear in mind the value of investigating these important constructs from a developmental perspective. In fact, a classic critique of much of the recent work in moral psychology has been the lack proper grounding in developmental theory. In both the case of Mikhail and Greene, it is fair to say that there have been few attempts at testing their theoretical claims developmentally. On this point, the work of Pellizzoni and colleagues deserves considerable praise; and we only hope that our work can add to their sizable contribution.
Contributions of Analytic Methods

Finally, we hope that the current paper becomes one of many more examples of the way that Bayesian methods can be used to solve a huge variety of problems. The models reported in this paper are not so complex that they could not have been handled by frequentist methods, but we elected to use a Bayesian approach both for philosophical reasons and also for the practical advantages inherent to this framework.
REFERENCES


Broeders, R., van den Bos, K., Müller, P. A., Ham, J. (2011). Should I save or should I not kill? How people sovle moral dilemmas depends on which rule is most accessible.


executive function across a wide age range: Children and adults find happy–sad more difficult than day–night. Developmental Science, 14, 481–489.


Main Track Demonstration Video

[Before starting the video] “Here is a train on the train tracks.” [Point to train]

"Have you ever seen a train before?"

“Do you know what sound a train makes?”

<table>
<thead>
<tr>
<th>child answers correctly</th>
<th>child answers incorrectly or doesn’t answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Yes, the train goes ‘Choo-choo!’”</td>
<td>“The train goes ‘choo-choo!’”</td>
</tr>
</tbody>
</table>

“Let’s see what happens when the train starts moving.”

[Play the video]

“Look, the train comes down the track like this!”

“Let’s look at that again!”

“Can you show me which way the train will go down the track?”

[Play the video again]
Switch Demonstration Video

[Before starting the video] “Here is another train on the train tracks.” [Point to train]

“What do you think this switch does?”

<table>
<thead>
<tr>
<th>child answers correctly</th>
<th>child answers incorrectly or doesn’t answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Yes, when someone flips this switch, the train goes down the side track!”</td>
<td>“When someone flips this switch, the train goes down the side track!”</td>
</tr>
</tbody>
</table>

“Let’s see how this switch works.”

[Play the video]

“Oh look, he flips the switch and the train goes down the side track!”

“Let’s look at that again! What do you think is going to happen if he flips the switch?”

[Play the video again]
“Ned” (*Loop Track* case)

“One day Ned is walking near train tracks. But suddenly, Ned sees a train that is out of control.” [Point to train.]

“Farther down the track there are 5 people on the track who will be hit by the train.” [Point out five people.]

“But Ned is next to a switch that can turn the train to the side track.” [Point to switch.]

“Do you remember what the switch does?”

<table>
<thead>
<tr>
<th>[child answers correctly]</th>
<th>[child answers incorrectly or doesn’t answer]</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Yes, the switch makes the train go down the side track”</td>
<td>“The switch makes the train go down the side track. So, instead of coming down this track [Point to main track.], the train would go around this track [Point to side track].”</td>
</tr>
</tbody>
</table>

“There is 1 man standing on the side track.” [Point to 1 man]

“If the train hits this man, that means it will *NOT* hit the 5 people on the track.”

“What do you think Ned should do?”

“Should Ned flip the switch, or should he *not* flip the switch?”

*[Wait for response. Repeat question and restate dilemma if necessary.]*
“Hank” *(Bystander case)*

“Hank is walking near train tracks one day. But then, Hank sees a train that is out of control.” *[Point to train]*

“Farther down the track are 5 people who will be hit.” *[Point to 5 people]*

Hank is next to a switch that can turn the train on to a side track. *[Point to switch.]*

“Do you remember what the switch does?”

<table>
<thead>
<tr>
<th>[child answers correctly]</th>
<th>[child answers incorrectly or doesn’t answer]</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Yes, The switch makes the train go down the side track.”</td>
<td>“The switch makes the train go down the side track. So, instead of coming down this track <em>[Point to main track.]</em>, the train would go around this track <em>[Point to side track.]</em>.</td>
</tr>
</tbody>
</table>

“But there is 1 man standing on the side track.” *[Point to 1 man]*

“If the train is turned on the side track, it will hit the one man. This would save the 5 people.”

“What do you think Hank should do?”

“Should Hank flip the switch, or should he *not* flip the switch?”

*[Wait for response. Repeat question and restate dilemma if necessary.]*
“Frank” (Footbridge case)

“Frank is on a bridge that passes over train tracks. But then, Frank sees a train that is out of control.” [Point to train]

“Farther down the track there are 5 people who will be hit.” [Point to 5 people]

“But next to Frank is a large man.” [Point to large man]

“Frank can push the man onto the track in the path of the train. This would slow down the train and save the 5 people on the track.”

“What do you think Frank should do?”

“Should Frank push the man onto the track, or should he not push the man?”

[Wait for response. Repeat question and restate dilemma if necessary.]
“Joe”  \( (\text{Bystander}_{\text{inv}} \text{ case}) \)

“Joe is walking near the train tracks one day. Suddenly, Joe sees a train that is out of control.” \([\text{Point to train.}]\)

“Farther down the track there is 1 person who will be hit. Joe is next to a switch. He can flip the switch to turn the train on to a side track.”

“Do you remember what the switch does?”

<table>
<thead>
<tr>
<th>[child answers correctly]</th>
<th>[child answers incorrectly or doesn’t answer]</th>
</tr>
</thead>
</table>
| “Yes, The switch makes the train go down the side track” | “The switch makes the train go down the side track. So, instead of coming down this track [\(\text{Point to main track.}\), the train would go around this track [\(\text{Point to side track.}\).]

“But there are 5 people standing on the side track.” \([\text{Point to train.}]\)

“If the train is turned on the side track, it will hit the 5 people. This would save the 1 person.”

“What do you think Joe should do?”

“Should Joe flip the switch, or should he not flip the switch?”

\([\text{Wait for response. Repeat question and restate dilemma if necessary.}]\)
“Bob” (*Bystander* 

absent case)

“Bob is walking near train tracks one day. But then, Bob sees a train that is out of control.” [*Point to train*]

“Farther down the track are 5 people who will be hit.” [*Point to 5 people*]

Bob is next to a switch that can turn the train on to a side track. [*Point to switch.*]

“Do you remember what the switch does?”

<table>
<thead>
<tr>
<th>child answers correctly</th>
<th>child answers incorrectly or doesn’t answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Yes, The switch makes the train go down the side track.”</td>
<td>“The switch makes the train go down the side track. So, instead of coming down this track [<em>Point to main track.</em>], the train would go around this track [<em>Point to side track.</em>].”</td>
</tr>
</tbody>
</table>

“There is no one standing on the side track.” [*Point to empty side track*]

“If the train is turned on the side track, it would save the 5 people.”

“What do you think Bob should do?”

“Should Bob flip the switch, or should he not flip the switch?”

[Wait for response. Repeat question and restate dilemma if necessary.]
Happy-Sad Task

Lagattuta et al., (2011)

*Using the 20 happy and sad cards.*

“Here is a picture of a face. Is it happy or sad? (*wait for participant’s response*)

“Right, happy!”

“Here is another picture of a face. Is it happy or sad?” (*wait for participant’s response*)

“Right, sad!”

“OK, we are going to play an opposite game”

“When I show you a picture of a happy face, I want you to say, 'SAD'! And, when I show you a picture of the sad face, I want you to say, 'HAPPY'!”

“So, let’s say the rules again. When I show you a happy face, you say _______.” (*participant responds*)

“And, when I show you a picture of the sad face, you say _______.” (*participant responds*)

“OK, let’s practice.”

*Show 1 card at a time. Place it face up on the table, and be sure camera can see.*

*The participant must get 4 practice trials correct in a row before moving on to the actual task.*

*If the participant responds incorrectly on a trial, prompt them by saying ‘Remember, this is an opposite game. When I show you a happy face, you say ‘Sad’. And when I show you a sad face, you say ‘Happy’.***

*After 4 consecutive successes, move to actual task.*
“OK, let’s play the opposite game!”

*Begin sequence of 20 cards.*
Receptive Vocabulary (PAL Test)

This task uses the 16-image sheet. The sheet must be within reaching distance of the participant, and within viewing distance of the camera.

“Now we’re going to play another game.”

“In this game, I’ll say the name of something and you point to the picture.”

“Are you ready?”

nurse
dentist
mittens
helicopter
sandals
kayak
trumpet
saxophone
Figure 15. Image used for PAL receptive vocabulary task. Participants were asked to point to the picture of the object that was named.
Contents False Belief Task

“Now we’re going to play a new game.”

[Pull out the crayon box]

“Here’s a crayon box. What do you think is inside the crayon box?”

[crayon box is opened]

“Let’s see…it’s really candles inside!”

[The Crayon box is closed]

“Okay, what is in the crayon box?”

[Then a toy figure is produced]

“Alvin has never ever seen inside this crayon box. Now here comes Alvin. He’s been asleep.”

“So, what does Alvin think is in the box? crayons or candles?”

“Did Alvin see inside this box?”

This task was derived from Perner and colleagues (1987) and Wellman and Liu (2004).
APPENDIX B

R code

```r
setwd('~/Rwd')

## packages used: lme4, rjags, modeest, ggmcmc, polycor, VIM
## sourced functions: disp_function, post_plot_function

source('plot_post_function.R')
source('disp_function.R')

## read in data

d_prnt <- read.csv('dissertation_data.csv', stringsAsFactors = FALSE)
d_child <- read.csv('social_minds_data.csv')
d_fam_id <- read.csv('diss_family_id.csv', header = FALSE)

## bookkeeping

N <- 129
d_child <- d_child[1:N, 1:37]
d_child$dilemma_ord <- as.numeric(d_child$dilemma_ord) - 1

## correcting data entry errors

d_prnt$birth[39] <- '09/28/10'
d_prnt$birth[75] <- '12/26/08'
d_prnt$birth[89] <- '03/17/10'
d_prnt$birth[110] <- '06/05/09'

###
## Exclude non-first trolley cases for first 9 participants
###
d_child$ned[c(1:2, 4:7)] <- NA
d_child$frank[c(2:6, 8:9)] <- NA
d_child$Shank[c(1, 3:5, 7:9)] <- NA

###
## Calculate age in months
```
## get rid of extra white space

d_prnt$birth <- gsub(" ", "", d_prnt$birth)

## write function to calculate age in months
CalcMonths <- function(date, birth){
## bookkeeping
  n <- length(date)
  date <- as.Date(date, format = "%m/%d/%Y")
  birth <- as.Date(birth, format = "%m/%d/%Y")

## some math
  dif_days <- date - birth
  age_m <- dif_days/30.416667

## cleaning up
  age_m <- as.vector(age_m)

  return(age_m)
}

age_m <- CalcMonths(d_prnt$start, d_prnt$birth)

## Define quick wrapper function for centering
Center <- function(x){
  x_center <- as.vector(scale(x, center = TRUE, scale = FALSE))

  return(x_center)
}

## function for identifying siblings and assigning a family ID
FindSibs <- function(x){
## bookkeeping
  n <- length(x)

## initialize vector to fill
  new_id <- rep(NA, n)
  new_id[1] <- 1

## loop over fam_id to find duplicates
for(i in 2:n){
    ## when finding duplicates
    if(x[i] == x[i-1]){  
        new_id[i] <- new_id[i-1]  
    }

    ## when not finding duplicates
    if(x[i] != x[i-1]){  
        new_id[i] <- 1+new_id[i-1]  
    }
}  
return(new_id)
}

###
# Compute Effortful Control subscale of CBQ
###
eff_cntrl <- apply(d_prnt[3:14], 1, sum)

###
# Compute Happy-Sad task score
###
hap_sad <- apply(d_chld[17:36], 1, sum)

###
# Compute receptive vocab score (PAL)
###
vocab <- apply(d_chld[9:16], 1, sum)

###
# Combine data
###
d <- cbind(d_chld[, c(1:8, 37)], d_prnt[, c(2, 15, 17:21)], age_m, eff_cntrl, hap_sad, vocab)

## preliminary
d$ToM <- ifelse(d$f.bel.memory == 1, d$f.bel.target, 0)

## fixing NAs on belief
miss <- which(is.na(d$f.bel.target))
d$ToM[miss] <- NA
## code utilitarianism

d$util <- ifelse(d$frank == 1 & d$ned == 1 & d$hank == 1, 1, 0)

################################################
# Identifying participants who conformed
# to the Contact Principle and the Doctrine
# of Double Effect.
################################################

contact <- rep(NA, N)
dde <- rep(NA, N)

## function to identify conformers
Conform <- function(dat, x0, x1, y){
  ## bookkeeping
  n <- nrow(dat)
  for(i in 1:n){
    ## exclude missing
    if(any(is.na(dat[, x0][i]), is.na(dat[, x1][i]))){
      y[i] <- NA
    } else{
      ## identify conformers
      if(dat[, x0][i] == 0 & dat[, x1][i] == 1){
        y[i] <- 1
      } else{
        ## identify non-conformers
        y[i] <- 0
      }
    }
  }
  return(y)
}

contact <- Conform(d, x0 = 'frank', x1 = 'ned', y = contact)
dde <- Conform(d, x0 = 'ned', x1 = 'hank', y = dde)

###
# Aggregate our data frames
###
d <- cbind(d, contact, dde, d_fam_id)
colnames(d)[25] <- 'fam_id'

###
# Excluding the too-old kid
###

d <- d[-57,]

###
# Exclude participants who failed controls
###

d2 <- d[which(d$joe == 0),]

###
# select "final" dataset and calculate missingness rate
###

d0 <- d2[, c('ID', 'new_famid', 'age_m', 'contact', 'dde', 'hap_sad', 'ToM', 'vocab', 'sex', 'educ',
           'income', 'bilingual')]

d0 <- as.data.frame(d0)

## calculate missingness
mean(is.na(d0[, c(-1, -2)]))

###
# Visualizing pattern of missingness
###

library(VIM)

matrixplot(d0[, c(-1, -2)], sortby = 'age_m', interactive = FALSE, ylab = NA, main = 'Summary of Missingness', bty = 'l')

###
# Correlations
###

print(round(cor(cbind(d2$vocab, d2$age_m, d2$hap_sad, d2$eff_cntrl, d2$educ), use = 'pairwise'), digits = 2))
library(rjags)

## Specify JAGS model
modelstring <- 'model{
    for(i in 1:N){
      x[i] ~ dbern(theta)
    }
    x_pred ~ dbinom(theta, N)
    theta ~ dbeta(1, 1)
}

## write model to .txt file
writeLines(modelstring, con = 'model.txt')

## set up data
dat <- list(
  x = d2$dde,
  N = nrow(d2)
)

## Run the model
model <- jags.model(file = 'model.txt', data = dat, n.chains = 3)
update(model, 1000)
samps <- coda.samples(model, c('theta', 'x_pred'), n.iter = 10000)

## Inspecting the posterior
disp(samps)
plot(samps, density = FALSE)

## Bind chains
samps_mat <- rbind(as.matrix(samps[[1]]), as.matrix(samps[[2]]), as.matrix(samps[[3]]))
plotPost(samps_mat[, 1], comp_val = 0.25, main = 'Prob. of Conforming to Doctrine of Double Effect')

###
# Bayesian regression models
###
library(modeest)

modelstring <- '
model{
  ## Level 1 Model
  for(i in 1:N){
    y[i] ~ dbern(pi[i])
    pi[i] <- 1/(1 + exp(-eta[i]))
    eta[i] <- beta[id[i]]
  }

  ## Level 2 Model
  for(j in 1:J){
    beta[j] ~ dnorm(mu_beta[j], tau_00)
    mu_beta[j] <- gam
  }

  ## priors on gamma and sigma squared
  gam ~ dnorm(0, 1.0E-3)
  sig_sq ~ dt(0, 1, 5) T(0, )

  ## transformations
  tau_00 <- 1/sig_sq
}
'

writeLines(modelstring, con = "model.txt")

## Setting up data and initial values
dat <- list(
  y = d2$dde,
  id = d2$new_famid,
  N = nrow(d2),
  J = max(d2$new_famid)
)

## Initialize model
m0 <- jags.model(file = 'model.txt', data = dat, n.chains = 3)
## Burn in
update(m0, 10000)

## Specify parameters to monitor, then run model
params <- c('gam', 'sig_sq')
m0_samps <- coda.samples(m0, params, n.iter = 50000, thin = 5)

## Examine posterior
disp(m0_samps)
plot(m0_samps)

## Diagnosing model
dic.samples(m0, n.iter = 5000)
gegelman.diag(m0_samps)

## compute ICC using mode of sigma_sq
tau_00 <- disp(m0_samps)['sig_sq', 'Mode']
icc <- tau_00/(tau_00 + (pi^2)/3)

## compute HPD for ICC
tau00_lo <- disp(m0_samps)['sig_sq', 'HPD low']
icc_lo <- tau00_lo/(tau00_lo + (pi^2)/3)
tau00_hi <- disp(m0_samps)['sig_sq', 'HPD high']
icc_hi <- tau00_hi/(tau00_hi + (pi^2)/3)

## Make some nice plots
S <- ggs(m0_samps)
ggs_density(S, family = 'p') + xlim(0, 10)
ggs_traceplot(S)

## bind chains of posterior
samps_mat <- rbind(as.matrix(m0_samps[[1]]), as.matrix(m0_samps[[2]]),
as.matrix(m0_samps[[3]]), as.matrix(m0_samps[[4]]))
plotPost(samps_mat[, 2], comp_val = NULL, xlab = expression(sigma))
modelstring <- 'model{
## Level 1 Model
for(i in 1:N){
y[i] ~ dbern(pi[i])
pi[i] <- 1/(1 + exp(-eta[i]))
eta[i] <- beta_0[id[i]] + beta[1]*x1[i] + beta[2]*x2[i] + beta[3]*x3[i]
}

## Level 2 Model
for(j in 1:J){
  beta_0[j] ~ dnorm(mu_beta[j], tau_00)
  mu_beta[j] <- gam
}

## priors on beta, gamma, and sigma squared
  gam ~ dnorm(0, 1.0E-6)
  for(k in 1:3){
    beta[k] ~ dnorm(0, 1.0E-6)
  }
  sig_sq ~ dt(0, 1, 5) T(0, )

## priors on missing data
for(i in 1:N){
  x2[i] ~ dnorm(x2_mu[i], 1.0E-3)
  x3[i] ~ dnorm(x3_mu[i], 1.0E-3)
  x4[i] ~ dnorm(x4_mu[i], 1.0E-3)

  x2_mu[i] <- b[1] + b[2]*x1[i] + b[3]*x4[i]
  x3_mu[i] <- b[4] + b[5]*x1[i] + b[6]*x4[i]
  x4_mu[i] <- b[7] + b[8]*x1[i]
}

## priors on b
for(k in 1:8){
  b[k] ~ dnorm(0, 1.0E-3)
}

## transformations
tau_00 <- 1/sig_sq
}'
writeLines(modelstring, con = "model.txt")

## Setting up data
dat <- list(y = d2$contact, id = d2$new_famid, x1 = Center(d2$sage_m), x2 = Center(d2$hap_sad), x3 = Center(d2$eff_cntrl), x4 = Center(d2$vocab), N = nrow(d2), J = max(d2$new_famid))

## Initialize the model
m1 <- jags.model(file = 'model.txt', data = dat, n.chains = 3)

## Burn in
update(m1, 1000)

## Specify parameters to monitor, then run model
params <- c('gam', 'beta', 'sig_sq')
m1_samps <- coda.samples(m1, params, n.iter = 50000, thin = 5)

## Examine posterior
disp(m1_samps, digit = 2, csv = TRUE)
plot(m1_samps, density = FALSE, col = c('navy', 'purple', 'skyblue'))

## Diagnosing model
dic.samples(m1, n.iter = 5000)
gelman.diag(m1_samps)

###
# Age predicting conforming to DDE
###

modelstring <- '
model{
## Level 1 Model
for(i in 1:N){
    y[i] ~ dbern(pi[i])
    pi[i] <- pt(eta[i], 0, 1, 3)
    eta[i] <- beta_0[id[i]] + beta[1]*x1[i]
}'}
## Level 2 Model

```
for(j in 1:J){
    beta_0[j] ~ dnorm(mu_beta[j], tau_00)

    mu_beta[j] <- gam
}

## priors on gamma, beta, and tau

gam ~ dt(0, 1, 3)
beta[1] ~ dt(0, 1, 3)
sig_sq ~ dt(0, 1, 3) T(0, )

## transformations

tau_00 <- 1/sig_sq
```

```r
writeLines(modelstring, con = "model.txt")
```

## Setting up data and initial values

```
dat <- list(
    y = d2$dde,
    id = d2$new_famid,
    x1 = Center(d2$age_m),
    N = nrow(d2),
    J = max(d2$new_famid)
)
```

## Initialize the model

```
m2 <- jags.model(file = 'model.txt', data = dat, n.chains = 3)
```

## Burn in

```
update(m2, 1000)
```

## Specify parameters and run the chains

```
params <- c('gam', 'beta', 'sig_sq')
m2_samps <- coda.samples(m2, params, n.iter = 50000, thin = 5)
```

## Examine posterior

```
disp(m2_samps, digit = 2, csv = TRUE)
plot(m2_samps, density = FALSE, col = c('navy', 'purple', 'skyblue'))
```

## Diagnosing model

```
dic.samples(m2, n.iter = 5000)
gelman.diag(m2_samps)
```
## Make nice plot of posterior

```r
samps_mat <- rbind(as.matrix(m2_samps[[1]]), as.matrix(m2_samps[[2]]), as.matrix(m2_samps[[3]]))

plotPost(samps_mat[, 1], comp_val = 0, xlab = expression(beta))
```

### # ToM predicting conforming to DDE ###

```r
modelstring <- '
model{
  ## Level 1 Model
  for(i in 1:N){
    y[i] ~ dbern(pi[i])
    pi[i] <- pt(eta[i], 0, 1, 1)
    eta[i] <- beta_0[id[i]] + beta[1]*x1[i]
  }

  ## Level 2 Model
  for(j in 1:J){
    beta_0[j] ~ dnorm(mu_beta[j], tau_00)
  }  
  mu_beta[j] <- gam

  ## priors on gamma, tau, and beta
  gam ~ dnorm(0, 1.0E-6)
  sig_sq ~ dt(0, 1, 5) T(0, )
  beta[1] ~ dnorm(0, 1.0E-6)

  ## priors on missing data
  for(i in 1:N){
    x1[i] ~ dbern(pi_x1[i])
    pi_x1[i] <- 1/(1 + exp(-eta_x1[i]))
    eta_x1[i] <- b[1] + b[2]*x2[i] + b[3]*x3[i]
    x2[i] ~ dnorm(mu_x2[i], 1.0E-6) T(0, )
    mu_x2[i] <- b[4] + b[5]*x3[i]
  }

  ## priors on b
  for(k in 1:5) {
    b[k] ~ dnorm(0, 1.0E-3)
  }
}
```
## transformations

tau_00 <- 1/sig_sq

```
writeLines(modelstring, con = "model.txt")
```

## Setting up data and initial values

dat <- list(y = d2$dde,
             id = d2$new_famid,
             x1 = d2$ToM,
             x2 = d2$vocab,
             x3 = d2$age_m,
             N = nrow(d2),
             J = max(d2$new_famid))

## Initialize the model

m3 <- jags.model(file = 'model.txt', data = dat, n.chains = 3)

## Burn in

update(m3, 1000)

## Specify parameters and run the chains

params <- c('gam', 'beta', 'sig_sq')
m3_samps <- coda.samples(m3, params, n.iter = 50000, thin = 5)

## Examine posterior

disp(m3_samps, digit = 2, csv = TRUE)
plot(m3_samps, density = FALSE, col = c('navy', 'purple', 'skyblue'))

## Diagnosing model

dic.samples(m3, n.iter = 5000)
gelman.diag(m3_samps)
### Age x ToM Interaction Model

```r
modelstring <-
model{
## Level 1 Model
for(i in 1:N){
y[i] ~ dbern(pi[i])
pi[i] <- pt(eta[i], 0, 1, 1)
eta[i] <- beta_0[id[i]] + beta[1]*x1[i] + beta[2]*x2[i] + beta[3]*x3[i]
}

## Level 2 Model
for(j in 1:J){
  beta_0[j] ~ dnorm(mu_beta[j], tau_00)
  mu_beta[j] <- gam
}

## priors on beta, gamma, and sigma squared
for(k in 1:3){
  beta[k] ~ dnorm(0, 1.0E-3)
}
gam ~ dnorm(0, 1.0E-3)
sig_sq ~ dt(0, 1, 5) T(0, )

## priors on missing data
for(i in 1:N){
x2[i] ~ dbern(pi_x2[i])
pi_x2[i] <- 1/(1 + exp(-eta_x2[i]))
eta_x2[i] <- b[1] + b[2]*x1[i] + b[3]*x3[i]
x3[i] ~ dnorm(mu_x3[i], 1.0E-3)
mu_x3[i] <- b[4] + b[5]*x1[i] + b[6]*x4[i]
x4[i] ~ dnorm(mu_x4[i], 1.0E-3)
mu_x4[i] <- b[7] + b[8]*x1[i]
}

## priors on b
for(k in 1:8){
  b[k] ~ dnorm(0, 1.0E-3)
}

## transformations
tau_00 <- 1/sig_sq
}

```
writeLines(modelstring, con = "model.txt")

## Setting up data
dat <- list(
  y = d2$dde,
  id = d2$new_famid,
  x1 = Center(d2$age_m),
  x2 = d2$ToM,
  x3 = Center(d2$age_m)*Center(d2$ToM),
  x4 = Center(d2$vocab),
  N = nrow(d2),
  J = max(d2$new_famid)
)

## Initialize the model
m4 <- jags.model(file = 'model.txt', data = dat, n.chains = 3)

## Burn in
update(m4, 1000)

## Specify parameters to monitor, then run model
params <- c('gam', 'beta', 'sig_sq')
m4_samps <- coda.samples(m4, params, n.iter = 150000, thin = 15)

## Examine posterior
disp(m4_samps, digit = 2, csv = TRUE)
plot(m4_samps, density = FALSE, col = c('navy', 'purple', 'skyblue'))

## Diagnosing model
dic.samples(m4, n.iter = 5000)
gelman.diag(m4_samps)