MORAL RULES AND REPRESENTATIONS: A DEVELOPMENTAL PERSPECTIVE

by

SYDNEY LEVINE

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Psychology

written under the direction of

Alan M. Leslie

and approved by

________________________

________________________

________________________

________________________

New Brunswick, New Jersey

October, 2016
ABSTRACT OF THE DISSERTATION

Moral Rules and Representations: A Developmental Perspective

By SYDNEY LEVINE

Dissertation Director:

Alan M. Leslie

If a grammar is taken to be a set of rules that operate over a series of representations (Chomsky, 1980), then this dissertation uses empirical developmental methods to begin to describe components of a moral grammar. Chapter 1 reviews the current state of the literature concerning moral judgment, with a particular focus on the role of intention in moral judgment and developmental implications. Chapter 2 proposes a “conversion rule” (a good intention prior) that translates an impoverished stimulus into a rich representation of a morally charged act, which is then suitable for the application of moral rules. In particular, the conversion rule proposes that when an agent’s action results in both a good and bad effect, the agent intends the good and not the bad effect. Chapter 3 argues that two moral rules used by preschoolers (the means principle and the implied consent principle) are written in abstract rather than concrete terms, taking inputs such as hierarchical action relations and utility structures rather than directly observable behaviors. Chapter 4 presents evidence that preschoolers and adults use a moral rule that draws on representations of the choices available to a moral agent. Chapter 5 draws together the chapters with concluding remarks.
Acknowledgements and Dedication

I became an undergraduate at Yale University in the Fall of 2005. On the first day of Introduction to Cognitive Science, Brian Scholl informed us that the mind was made of mental organs called “modules.” I spent that semester reading selections from Fodor, Pylyshyn, Gallistel, and Leslie. Six years later, coming to Rutgers to pursue a doctoral degree in cognitive science felt like a pilgrimage to the source of all the great ideas. Studying in this department has been an extraordinary honor that I am immensely grateful for.

Primary thanks are due to my adviser and dissertation committee chair, Alan Leslie. Through long conversations over the past five years, I have learned about experimental design and execution, the role of developmental science in the field of cognitive science at large, what it means to be a scientist pursuing the truth, and innumerable details of Scottish history. Without his encouragement, I would likely be going on to take a job as a podiatrist or prison warden at this juncture.

I also owe an immense debt to John Mikhail. Working carefully through his book, *Elements of Moral Cognition*, taught me how to think rigorously about cognitive science and showed me what the best version of theory of cognitive science could look like. I visited John at Georgetown Law School in the summer of 2012, and the discussions we had during that time defined the research program I worked on throughout graduate school and will likely continue to shape my interests for years to come.
During the course of my graduate work, I was also mentored by Stephen Stich and Edouard Machery, whose role in bringing empirical work to bear on philosophical questions I deeply admire and seek to emulate.

I am also grateful for the guidance of many members of the Rutgers Center for Cognitive Science faculty: the other members of my committee, Jacob Feldman and Julien Musolino, in addition to Randy Gallistel, Rochel Gelman, Gretchen Chapman, and Alvin Goldman.

Among my colleagues at Rutgers, I owe particular thanks to Katya Saunders, whose work laid the groundwork for many of the studies I conducted, and my officemate, Lu Wang, whose brilliance as a thinker and scientist I have admired since we met and who was always willing to drop what she was doing to help me puzzle through an idea. I was exceedingly fortunate to journey through graduate school with Nora Isacoff and her “idiosyncratic” dog Daisy (ך”ל). In addition to her invaluable advice about pedagogy and our ongoing conversation about the big issues in cognitive science and academia, our friendship was a defining force throughout my graduate school years.

Special thanks are also due to my colleagues Michelle Cheng, Gala Stojnic, Alex D’Esterre, Talia Walzer, Daglar Tanrikulu and the members of the Cognitive Development Lab.

I am also grateful for being welcomed into the community of philosophy graduate students at Rutgers and encouraged to join their unending, frequently beer-laden conversation. In particular, Lisa Miracchi, Will Fleisher, Anton Johnson, David Rose, Justin Sharber, and Georgi Gardiner were cherished colleagues and confidants.
I am indebted to Josh Tenenbaum and the computational cognitive science group at MIT (especially Max Kleinman-Weiner, Tobias Gerstenberg, and Chris Baker), who welcomed me to visit their lab in the final year of my graduate studies and whose work feels like it pushes the field in the right directions. The invigorating conversations I had with Josh in my fourth year were some of the highlights of graduate school and helped me understand my own work more completely and how it fits into the cognitive science tradition.

I am equally grateful to Fiery Cushman and the moral psychology research group at Harvard (especially Adam Morris, who specifically asked to be put in the acknowledgments, Jonathan Phillips, Wouter Kool, and Justin Martin). Somehow, Fiery managed to read and provide incredibly insightful comments on nearly all of my manuscripts even before we were working together. I am grateful for his clear-headed mentorship.

In many ways this work emerged out of a conversation I began in the mid-2000s with two intellectual partners, Jacob Abolafia and Raffi Magarik, on the nature of the good life and on what matters. That conversation continues to this day and if this work has managed to inform it at all, I will consider my graduate school career a success.

Thanks to the friends who supported me during this time: Anna Shneidman, Michael Weingart, Tanya Farber, Reza Green, Moshe Silver, Gary Rendsberg, Beth Kramer, Chani Kott, Rachel Lemonik, Yonah Lemonik, Jessica “JJ” Weiss, Arianna Schudrich, Benjamin Boyd, Viki Bedo, Josh Rottman, and Victor Kumar.
Finally, thanks to my family, Karen Levine, Samantha Levine, Noah Levine, and Jess Sarli, who endured great bouts of anxiety and existential angst all throughout but especially in the final stages of this dissertation. They haven’t cast me out yet, clear testament to their sainthood.

This dissertation is dedicated to the memory of my father, Dennis Levine, who is still, somehow, cheering the loudest.
**Table of Contents**

Abstract .............................................................................................................................................. ii

Acknowledgements and Dedication ................................................................................................. iii

Table of Contents................................................................................................................................. iv

List of Tables ........................................................................................................................................ v

List of Figures ....................................................................................................................................... vi

I. Introduction ....................................................................................................................................... 1

II. Presumed Innocent? How Tacit Assumptions of Intentional Structure Shape Moral
Judgment .............................................................................................................................................. 23

III. Preschoolers’ Use of Moral Rules and Representations ................................................................. 53

IV. The Role of Choice in Moral Judgment ......................................................................................... 80

V. Conclusion ...................................................................................................................................... 101

Appendix A: Pink Scale Training ......................................................................................................... 104

Appendix B: Stimuli for Chapter II Studies ......................................................................................... 107

Appendix C: Stimuli for Chapter III Studies ......................................................................................... 117

Appendix D: Stimuli for Chapter IV Studies ......................................................................................... 128

References ........................................................................................................................................ 147
List of Tables

Table 2.1: Correlation matrix for should, Likert rating, and intention questions ..........46
List of Figures

Figure 2.1: Two intention structures are compatible with one causal structure........31
Figure 2.2. The relationship between the causal structure and the intention structures of the story in Experiment 1 .................................................................35
Figure 2.3. Adult subjects’ deontic judgments.................................................39
Figure 2.4: Adult subjects’ judgments of the intention of the agent..................40
Figure 2.5: Preschool subjects’ moral judgment of the action of the agent.........43
Figure 2.6: Preschool subjects’ Likert ratings of the action of the agent ..........44
Figure 2.7: Preschool subjects’ judgments of the intention of the agent..........45
Figure 3.1: Selected scenes from Side-Effect and Means Conditions (Experiment 1) and Means without Stealing (Experiment 1b) .................................................................56
Figure 3.2: Subjects’ responses to the question, “Should she have done that?”.....58
Figure 3.3: Subjects’ ratings of the action of the protagonist............................59
Figure 3.4: Subjects’ responses to the question, “Should she have done that?” from Study 1 (first two bars) and Study 1b (last bar) .................................................................60
Figure 3.5: Subjects’ ratings of the action of the protagonist Experiment 1 (first two bars) and Experiment 1b (last bar). Error bars shows standard error of the mean........61
Figure 3.6: Selected scenes from Side-effect and Means condition of Experiment 2 ....63
Figure 3.7: Subjects’ responses to the question, “Should he have done that?”..........64
Figure 3.8: Subjects’ ratings of the action of the protagonist............................65
Figure 3.9: Subjects’ responses to the question, “Did Billy squish the apple on purpose?”......................................................................................................................66
Figure 3.10: Scenes from the Implied Consent Conditions of Experiment 3 and 4 ......68
Figure 3.11: Subjects’ responses to the question, “Should she have done that?” ........69
Figure 3.12: Subjects’ ratings of the action of the protagonist. .........................70
Figure 3.13: Subjects’ responses to the question, “Should she have done that?” ........71
Figure 3.14: Subjects’ ratings of the action of the protagonist. .........................72
Figure 3.15: Subjects’ responses to the question, “Should she have done that?” ........74
Figure 3.16: Subjects’ ratings of the action of the protagonist. .........................75
Figure 3.17: Subjects’ responses to the question, “Should she have done that?” ........76
Figure 3.18: Subjects’ ratings of the action of the protagonist. .........................77
Figure 4.1: Selected scenes from each condition. ...........................................83
Figure 4.2: Subjects’ responses to the question, “Should she have done that?” ........86
Figure 4.3: Subjects’ ratings of the action of the protagonist. .........................87
Figure 4.4: Subjects’ responses to the question, “Did she make the one kid happy/sad on purpose?” .................................................................88
Figure 4.5: Percentage of subjects responding “yes” to the should question...........89
Figure 4.6: Subjects’ responses to the question, “Should she have done that?” ........91
Figure 4.7: Subjects’ ratings of the action of the protagonist. .........................91
Figure 4.8: Subjects’ responses to the question, “Did she make the one kid happy/sad on purpose?” .................................................................92
Figure 4.9: Subjects’ responses to the question, “Should she have done that?” ........94
Figure 4.10: Subjects’ ratings of the action of the protagonist. .......................95
Figure 4.11: Subjects’ responses to the question, “Did she make the one kid happy/sad on purpose?” .................................................................96
Figure 4.12: Subjects’ responses to the question, “Did Sally make all the kids on the right happy on purpose?’

Figure 4.13: Percentage of subjects responding “yes” to the should question.
I. Introduction: The Role of Intention in Moral Judgment: A Developmental Approach

In this review I will discuss the major theories of moral judgment (Section 1) and then turn to the role of intention in moral judgment (Section 2). Throughout, I will focus on developmental questions and implications.

1. Theories of Moral Judgment

In this section I will review five major theories of moral judgment, ending each section by noting the theory’s suggestion of how the capacity for moral judgment could develop in young children.

1.1. Moral Grammar

An analogy between moral cognition and linguistic cognition has been fruitfully developed by several authors (Dwyer, 2009; Mikhail, 2007, 2011; Harman, 2000; Rawls, 1971). I will focus on the most worked out of these views, Mikhail’s Moral Grammar Hypothesis. The Moral Grammar Hypothesis suggests that deontic judgments are the output of a modular system that runs a series of computations over highly structured, informationally rich mental representations. How do these structured, rich mental representations get formed from the impoverished stimuli available in the environment? Mikhail (2007, 2011) has suggested that a series of computational conversion translate the stimulus into a representation of the appropriate form – namely, one that can act as the input to the deontic computation.

In particular, the following series of conversion rules that operate on a stimulus (for example, a verbal or written narrative prompt) have been proposed. (1) Action descriptions (or act-tokens) are identified. (2) The temporal sequence of action descriptions is determined. (3) Causal dependencies of action descriptions are determined. (4) Valence of the action descriptions (moral goodness and badness) is determined. (5) Intentional structure of the actor is determined,
allowing for denotation of ends, means, and side-effects. (6) Preliminary deontic assessment of morally salient acts is determined and these acts are situated at locations in the structured representation.

The output of steps (1)-(6) would be a representation that makes explicit (in Marr’s (1982) sense) the information necessary to enter the computational process of deontic assessment (in which more complex moral grammatical rules are applied than those computed in step (6) – such as the Doctrine of Double Effect).

Mikhail takes a strong nativist approach to moral knowledge, pointing to the fact that a wealth of empirical evidence suggests that children are “intuitive lawyers”, seeming to know complex legal principles without explicit instruction, such as the role of intent in moral judgment (e.g., Cushman, Sheketoff, Wharton & Carey, 2013) and that false factual beliefs but not false moral beliefs can exculpate a moral agent (Chandler, et al, 2000). Mikhail also cites the fact that adults use abstract moral principles to make moral judgments that are not accessible to them upon introspection (Hauser et al, 2007; Cushman et al, 2006), suggesting that explicit learning has not necessarily played a major role in the development of their moral cognition.

1.2. Dual Process Views

Several authors have suggested the moral judgment is best understood through the lens of dual process cognition (Cushman, 2013, 2015; Crockett, 2013; Greene, 2013). Dual process approaches posit that both “fast” and “slow” systems are at work in a given cognitive task (for a review, see Evans, 2003). System I is typically characterized as automatic, fast, intuitive, non-reflective, and inflexible. This system is often described as being useful for its speed, though its major downside is the inability to be penetrated by contextual factors that are mediated by higher level cognitive processes. System II is characterized by careful, often conscious, rational deliberation typically associated with
reasoning. System II is computationally intensive and therefore slow, but its main benefit is its flexible ability to deal with changing circumstances.

This conflict between automatic and controlled processes seems to have hallmarks in the domain of moral judgment. This was first noticed by Greene (2001, 2013) who argued for a dual-process approach to moral judgment from behavioral responses and neuroscientific data about a now-infamous philosophical thought experiment: the trolley problem. There are two classic trolley problem cases. In the switch dilemma, a train is out of control and about to run over five people standing on the tracks. A bystander can flip a switch diverting the train to a sidetrack where only one person is standing. In the footbridge dilemma, a train is out of control and about to run over five people standing on the tracks. A bystander can throw a large person off a footbridge overlooking the tracks, causing the train to stop upon contact with the man. Typically, subjects permit acting in the former case but not the latter (e.g., Greene, 2001, 2009; Mikhail, 2007). Greene argues that in both cases the utilitarian rule “save as many lives as possible” is operative through System II cognition. However, in the footbridge case, an additional System I, emotionally driven response overrides the System II response and generates the judgment that the action is impermissible (Greene 2001; 2013). Greene finds neuroscientific evidence to support this claim, specifically that “emotional” regions of the brain show more activation in footbridge-style cases (Greene, 2001) and that patients with neurological damage to emotion regions make more utilitarian judgments in cases of these sorts (Koenigs, Young, et al., 2007).

Two recent theories (Cushman, 2013; Crockett, 2013) suggest that the dual-process approach to moral judgment can be instantiated by “model-free” and a “model-based”
reinforcement learning systems, which can be mapped onto action- and outcome-based value representations, respectively. Drawing on recent developments in hierarchical reinforcement learning models, these theories show how the two systems can be integrated to allow intention to be represented as hierarchical superordinate/subordinate-goal planning. Value can then be applied to the representations to yield moral judgments.

Greene suggests that moral judgment evolved as evolution’s way of solving the free-rider problem. That is, in many cases that arise in situations of cooperative living, it is beneficial for everyone if there is some communal resource that each person contributed to. However, it is beneficial for each individual to defect from cooperating and to reap the benefits of everyone else’s contribution (Greene, 2013). Of course, if everyone defected, there would be no communal resource, leaving everyone worse off. Morality is our way of heading off this problem before it begins. Given this evolutionary story, Greene clearly thinks that innate underpinnings of moral judgment are critical; though he stops short of describing specifically what he think is built in and what could be acquired through reliable cultural learning (but cf. Greene et al., 2004 for discussion of innateness of aversion to “personal force” scenarios).

Cushman & Miller (2013) provide more of an insight into how a piece of the dual process system for moral judgment might develop. They start out by pointing out that a negative emotional reaction to a morally charged stimulus might arise from two sources: outcome aversion and action aversion. Outcome aversion is the negative reaction a moral actor has when faced with the prospect that his action will harm a particular victim; it is an aversion to the suffering of another, possibly through the mechanism of empathy. Action aversion, on the other hand, is the negative response felt by the moral actor when
faced with performing a certain action. (Cushman & Miller suggest that third-party moral judgment may work partially via evaluative simulation of the actions of others.) The developmental insight is that action aversion could be acquired either through experience, observation, or imagination and could begin at a young age and continue throughout life.

1.3. Moral Rules

The Sentimental Rules Theory (Nichols 2002, 2004) posits that moral judgment is built on the combination of a Normative Theory—a set of rules prohibiting certain behaviors—and an affective mechanism—which leads to an emotional response when one of the rules are transgressed. The theory makes the distinction between a moral judgment, arising from the violation of one of the affectively backed norms, and a conventional judgment, the acknowledgment that a norm has been violated but without the emotional reactive component. The Sentimental Rules Theory suggests that transgressions that elicit the moral norm response-pattern are considered to be more serious, less permissible, and less authority-dependent than those that elicit the conventional response-pattern.

Are moral rules actually necessary, though, or can moral judgment be explained as an affective response to a stimulus? Nichols & Mallon (2006) used non-emotionally charged moral violations (a child breaking teacups) and artificially established rules (“don’t break any teacups”) to probe the role of rules in moral violations. They found that when a teacup was intended to be broken (as a means), subjects judged that the rule was violated whereas when the teacup was broken as a side-effect of an action, subjects were less likely to view the rule as being violated. This pattern of findings suggests that
the means/side-effect distinction in the moral domain could be rule-driven. The authors also are right to point out that the rules that are driving moral judgment may not be consciously accessible and instead could be generating moral intuitions that have been observed in a wide range of experiments (Mallon & Nichols, 2011).

Nichols, Kumar & Lopez (in press) propose that moral rules could be acquired by young learners, despite the fact that, in one analysis of a section of CHILDES, less than 5% of morally-relevant child-directed language concerned rules (Wright and Bartsch, 2008). Even without this empirical evidence, it is hard to believe that direct instruction could account for the sensitivity with which children apply rules subtly and to highly specific contexts (Mikhail, 2011; Levine & Leslie, 2015). If children are not learning rules didactically, what mechanism is available to them? Nichols and colleagues (in press) argue that if children have a set hypothesis space of moral rules (perhaps innate), then Baysian learning strategies, in particular the “size principle”, could allow children to learn that rules tend to prohibit intended consequences (that is, the rules are narrow in scope) rather than foreseen but unintended consequences (which would be wider in scope).

1.4. Social Domain Theory

Proponents of Social Domain Theory (eg: Nucci & Nucci, 1982; Turiel, 1998; Killen & Smetana, 2008) suggest that moral norms are psychologically distinct from non-moral norms, typically referred to as conventional norms. In particular, moral norms are those that are universalizable (apply to all people in all times), highly serious, and not authority contingent (no authority figure can appeal them). Conventional norms may be specific to a time and place, are typically less serious, and can be appealed by the relevant authority.
In addition to this psychological profile that goes along with moral norms, moral norms have a specific content: they concern justice, welfare, and rights.

Proponents of Social Domain Theory have a specific developmental hypothesis (Killen & Smetana, 2008). They predict that by 2.5 years old, children will distinguish between moral and conventional norms and respond to their violations differently (Smetana, 1981). Moreover, the psychological response to and conception of moral norms should be the same across cultures, regardless of environmental influences. In contrast, conventional norms emerge as children learn from those around them.

1.5. Moral Foundations Theory

The Moral Foundations Theory (Haidt, 2012; Haidt & Joseph, 2007; Graham, Haidt & Nosek, 2009) suggests that the moral domain is wider than proponents of Social Domain Theory propose. According to Moral Foundations Theory, every culture’s morality is built upon (at least) five moral foundations: Harm/care, Fairness/reciprocity, Ingroup/loyalty, Authority/respect, and Purity/sanctity (Haidt & Joseph, 2007). Each foundation has a distinct set of emotions that arise in response to violations that occur in that domain. Each foundation is supposed to have evolved to solve a unique problem that arose in the evolutionary landscape and therefore can be matched with a set of adaptive triggers (the things the foundation was “designed” to respond to) which is smaller than the actual domain (the things the foundation, in today’s world, actually responds to). For example, the trademark emotion attached to the Purity/sanctity foundation is disgust. The foundation arose as a way of avoiding microbes and parasites (Haidt & Joseph, 2007; Haidt, 2012)
and therefore is triggered as a response to waste products and diseased people (proper domain) as well as taboo ideas (actual domain).

While each culture’s morality is composed of five foundations, each culture may place different amounts of emphasis on each foundation. For example, political liberals emphasize the two “individualizing” foundations – harm and fairness – while political conservatives seem to place less relative importance on those foundations and place more weight on the other three “binding” foundations – ingroup, authority, and purity (Graham, Haidt, & Nosek, 2009).

Moral foundations theory emphasizes both innate and learned aspects of moral judgment. Each child is born endowed by evolution with informational traces of information concerning the five foundations. However, each child is tasked with the challenge of acquiring the particular moral culture of its surroundings.

2. Role of Intention in Moral Judgment

In many theories of moral judgment, intention plays an important role. In this section I will first point out that each theory makes a specific, untested assumption about intention inference in the moral domain. Then, I will review experimental data that speaks to the role of intention in the moral judgments of adults and preschoolers. Finally, I will ask if there is an analogous reliance on intention for more primitive socio-moral evaluation in infants.

2.1. Intention inference in the moral domain

Agents move through the world constantly starting causal sequences of events that can be parsed in infinite ways. For example, I raise my hand signaling that I want to
answer a question, but at the same time catch the eye of a student sitting behind me, whack someone who suddenly got up from her chair, create a small breeze, and increase the number of hands raised by one. Most onlookers would immediately infer that the former of these actions was the one that I intended, yet, on the face of things, it seems that the bare evidence supports inferences that any of them may have been the agent’s goal. How do we rapidly, automatically, and reliably determine the goal of agents, given the vast number of effects every action causes?

The problem of determining which effects an agent intends is particularly central to our capacity to make moral judgments. We have already established that a single action brings about a myriad of effects. The problem becomes critical in the moral domain when some of the effects of an action are morally good and some are morally bad. Moral judgments often hinge on subtle differences in the mental states of moral actors with respect to the morally good and bad effects (reviewed more in-depth below, but see Young, Cushman, Hauser & Saxe, 2007; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; Hamlin, Ullman, Tenenbaum, Goodman & Baker, 2013). Therefore, an on-looker must first determine which effects were intended before making a moral judgment. For example, if I raised my hand in order to whack the person who suddenly got up from her chair, I may be judged more morally culpable for the harm I caused her than if my whacking of her was accidental and my intended goal was to signal my eagerness to answer a question.

Current theories of intention inference (eg: Woodward, 2013; Gergely & Csibra, 2003; Biro & Leslie, 2007; Meltzoff, 2005; Premack, 1990) fall short of being able to explain how we infer the goal of a novel action (see Levine, Mikhail & Leslie, in prep).
Therefore, theories of moral cognition cannot simply hope to “plug in” a theory of intention inference as a solution to this problem for moral judgment. The solution to the problem of goal inference for a novel action may involve contextual knowledge that is specific to a particular domain. Put another way, domain-specific information may be critical to solving the problem of goal inference from novel actions. There may be a solution to the problem of goal inference for a novel action that is particular to the moral domain, has explicit moral content, and should therefore be appended to theories of moral cognition (Levine, Mikhail & Leslie, in prep).

While many theories of moral cognition that emphasize the role of intention in moral judgment simply assume that intention is inferred and begin their theoretical work there (eg: Young & Saxe, 2011; Nichols & Mallon, 2006), others are attempting to provide explanations of the way that the necessary action representations are built so that moral judgment can proceed. The theories in this latter category, while not always acknowledging the difficult nature of the problem of intention inference for novel actions, seem to have either tacitly or explicitly assumed a particular, untested assumption about intention inference. Namely, that if an agent causes an action that results in both morally good and morally bad effect, then the agent intends the good and not the bad effect. Call this the “good intention prior.” Below I will point out where this assumption comes up in three theories of moral cognition.

As mentioned above, the Moral Grammar Hypothesis (Mikhail, 2007; 2011) suggests that deontic judgments are the output of a modular system that runs a series of computations over highly structured, informationally rich mental representations. We construct these representations using a series of cognitive conversion rules that translate
the stimulus (a perception of a morally charged act) into a mental representation. Mikhail (2007; 2011) notes that constructing the structure of the intention of an agent given only the input available from previous conversion rules (e.g., temporal alignment of effects, causal structure) requires making an assumption about which of the morally good or bad effects in the causal sequence of events count as the goal of the actor. Mikhail’s suggestion is that we do this by assuming that the agent intends the good effects and not the bad effects of her action: the good intention prior.

While no other theory has made this explicit, several seem to tacitly assume a cognitive constraint like the good intention prior. For example, Greene (2013) suggests that we have a modular system that inspects the action plans of agents as part of the mechanism of moral judgment. The module is “myopic” in the sense that it is blind to harmful side-effects. This explains, among other things, why we issue harsher moral judgments to harm caused as a means than when similar harm is caused as a side-effect. Harm as a means is represented as a necessary part of the main branch of an action plan leading to the goal; this is seen by the module. Harm as a side-effect is represented on a side-branch of an action plan, outside the scope of the module. Despite his emphasis on the importance of processing the action plans of agents, Greene does not provide an account of how we infer which effects count as side-effects, means, and goals. However, in each of the action plans that he draws to illustrate how we represent the minds of moral agents in particular moral dilemmas, he indicates that saving lives is the goal of the agent. In the background of Greene’s theory, therefore, is the critical, untested assumption that we assume that agents intend the good effects and not the bad effects.
In addition, the two recent theories that develop the dual-process model in terms of “model-free” and “model-based” reinforcement learning systems (Cushman, 2013; Crockett, 2013), also seem to rely on the “good intention prior” as an assumption about intention inference. As discussed above, these theories show how model-based and model-free systems can integrate value to create goal/sub-goal hierarchical structures which can allow us to understand goal-directed action and planning. However, in order for this process to even get off the ground, an assumption exists in the background that the agent acts “out of concern for others rather than malice” (Cushman, 2013, pg. 283). Given this assumption, goals and sub-goals can be assigned to the agent and moral cognition can proceed. Again, that that assumption is empirically warranted remains an open question.

2.2. The role of intention in adult moral judgment

Despite the fact that theories of moral cognition are still struggling to answer the question of how intention is inferred in the moral domain (as reviewed above), there is significant work suggesting that once intention is inferred, it plays an important role in moral judgment.

The role of intention in moral judgment is typically studied by contrasting four kinds of cases: attempted harm (bad intention, neutral outcome), accidental harm (neutral intention, bad outcome), intentional harm (bad intention, bad outcome), and harmless actions (neutral intention, neutral outcome). For example, an accidental harm may involve tripping over a branch and pushing someone over whereas an intended harm may involve trying to push someone over and instead tripping over a branch. On this view, intention is seen as binary: either an action was done intentionally or it was done
accidentally. As Cushman rightly points out (2015), this is a somewhat impoverished way of thinking about the role of intention in moral judgment. An intention is better understood as an agent’s action plan, driven by his beliefs and desires, aimed towards bringing about a goal (eg, Bratman, 1987). When viewed this way, an intention can helpfully be decomposed into at least four elements: (1) the beliefs and desires of the agent, (2) the action plan of the agent (3) the agent’s agents, (4) the outcome (Cushman, 2015). Cushman points out that it is particularly useful to think about intention in terms of these “decomposed” elements because different features of moral judgment seem to map onto them. For example, moral character assessments are based largely on information about an agent’s desires and beliefs. Wrongness/permissibility judgments on information about planning, intentional action, and goals. Punishment and blame hinges largely on whether the agent was causally responsible for the outcome and the magnitude of the outcome. Each “element” of intention plays a unique role in each feature of moral judgment.

For simplicity’s sake, I will focus on the role of intention in moral permissibility/wrongness judgments, which hinge on an agent’s action plan. In one of the most stark demonstrations of the role of intention in moral judgment, Young and colleagues (2007) used a 2x2 design, crossing bad/neutral outcomes with bad/neutral intentions. The authors found main effects of outcome and intention; bad outcomes (regardless of intention) are judged more harshly than neutral outcomes and bad intentions (regardless of outcome) are judged more harshly than neutral intentions. Interestingly, there was an intention x outcome interaction showing a significant difference in judgments of neutral and negative outcomes when the actor had a neutral intention but
no difference between neutral and negative outcomes when the actor had a negative intention. Put another way, a negative intention demands just as harsh a judgment regardless of outcome. This finding, that intentional harm is judged worse than accidents and that intentional harm and attempted harm are judged similarly, has been investigated in various other settings as well and is by now well-established (e.g., Cushman, 2008; Cushman & Young, 2011; Malle & Knobe 1997; Knobe, 2005).

The canonical nature of this finding has recently been challenged in two different ways: first, by studies in moral domains other than harm. Second, by cross-cultural work.

Young & Saxe (2011) looked at the role of intention in purity violations by comparing cases of accidental, attempted, and intentional incest and assault. They found that intent played a much smaller role for evaluating purity violations than for harm violations. In particular, guilty intentions increase condemnation of harm violations to a greater extent than for purity violations. Innocent intentions seem to exculpate to a greater extent for harm violations than for purity violations. The authors conclude that, in the purity domain, moral judgment hinges primarily on outcome rather than mental state.

In addition, recent cross-cultural work suggests that the importance of intent for moral judgment (even in the harm domain) may be limited to the populations that have been extensively studied by psychologists, namely those living in western, educated, industrialized, rich, democratic (WEIRD) societies (see Henrich, Heine, Norenzayan, 2010 for a review of the psychological “truths” that turn out to be particular to WEIRD people). In a surprising recent finding, Barrett and colleagues show that across eight traditional, small-scale societies there is significant variation in the role of intention in moral judgment (2016). In fact, in the WEIRD population (participants tested in Los
Angeles), intent mattered the most for moral judgment. In one society (the Yasawa in Fiji) intent was essentially meaningless; participants judged accidents with equal severity as intentional harm. Between the western participants the Fijians, a huge range of reliance on intent was observed. This finding calls into question whether the wealth of studies that find a prominent role for intention in moral judgment truly reflect what is psychologically core, or something that is instead more contingent on environmental inputs.

2.3. The role of intention in preschooer moral judgment

Is the use of intention in moral judgment learned by living in a Western society, or is it party of core moral cognition? One way of answering this question is to investigate if very young children seem to rely on intention when making moral judgments.

Piaget (1965/1932) noticed that younger children tend to rely more on outcome than intent when deciding whether an action was wrong (therefore condemning accidents harshly), but then eventually reverse their judgments as they get older, instead relying on intent more than outcome (tending to excuse accidents and condemn attempts). Recent work contrasting accidental and intentional harm have continued to confirm these findings, locating the outcome-to-intent shift around the age of 4-5 years old (Cushman, Sheketoff, Wharton & Carey, 2013; Baird & Astington, 2004; Killen, Mulvey, Richardson & Jampol, 2011).

While it is now well-established that preschoolers use intent in their moral judgments of accidental and intentional harm, it is less well known if preschoolers use more subtle differences in intention to modulate their moral judgments.
Mikhail (2002) asked children for their moral judgments of an intentional battery case and a foreseeable battery case. In the intentional battery case, children were told about a doctor who cuts up one healthy person, harvests his organs, and manages to save the lives of five sick people, who would otherwise have died. In the foreseeable battery case, children were told a story about a person who is driving a train that has gone out of control; the train driver has the option of running over five people on the train track, or diverting the train onto a sidetrack where only one person will be run over.

In each case, the outcome is kept constant – that is, one person dies and five are saved. However, the battery (and subsequent homicide) of the moral patient in the organ-donation case is intentional – it is a necessary part of the action plan of the moral agent and a causally necessary requirement for the plan to be carried out successfully. On the other hand, in the train-driver case, the battery (and homicide) of the victim on the sidetrack are foreseen, but not a causally integral part of the agent’s action plan. The author’s question in this study was whether 8-12 year old children would be sensitive to this distinction. Indeed, children were more likely to call the organ-donation case “wrong” than the train-driver case. These cases are not controlled in the most ideal way; they vary on many dimensions besides the intentional status of the battery committed by the moral agent. However, this study provides preliminary support for the notion that even young children distinguish between intentional and foreseen battery.

Pellizzoni, Siegal, and Surian (2010) constructed a child-appropriate version of the original Sidetrack and Footbridge versions of the Trolley problem. Children (ages 3-5) were shown a set-up of Legos in which a ball was rolling towards five Lego people who would be knocked down if another Lego person did nothing. In the sidetrack version, the
Lego bystander could choose to pull a chord to redirect the ball onto a sidetrack, knocking over one Lego person. In the footbridge version, the Lego person could choose to knock a larger Lego person onto the track in front of the ball, thereby knocking over the one large person, but saving the five. Children in the sidetrack version were more likely to think that it was “right” for the Lego moral agent to act than children in the footbridge case. This is some evidence that children are distinguishing harm as a means in comparison to harm caused as a side-effect. Pellizzoni et al’s study is more carefully controlled than Mikhail (2002), though it still confounds physical contact with intentional battery. That is, a plausible alternative explanation of the author’s findings is that children are operating based on a “contact principle” – physically contacting another agent is impermissible – rather than a more sophisticated assessment of intentional harm version foreseeable harm.

Saunders et al. (2014) used a novel trolley-like paradigm which avoids this concern. Children (ages 3-5) were told a story about a squirrel, which likes to eat children’s food. There are children in the park who are in danger of losing their cookies to the squirrel unless a moral agent, who is witnessing the scene, acts. In the side-effect version, the moral agent puts up a gate and re-directs the squirrel so that the squirrel only eats one child’s cookie rather than five children’s cookies. In the intentional harm condition, the moral agent takes a cookie from one child and feeds it to the squirrel so that the other children can eat their cookies in the meantime. These stories vary whether the harm to the moral patient was intentional or merely foreseeable and intentionality of the harm is not confounded with contact. In this study, children are much more likely to say that the moral agent should act in the side-effect condition than in the intentional harm condition.
This is stronger evidence that even very young children have moral intuitions that express the doctrine of double effect. However, much work remains to be done to support these claims.

2.4. The role of intention in socio-moral evaluation in infants

There is now a significant body of evidence that by 6 months old, infants view the actions of agents as goal-directed (Woodward, 2013). Early studies of this phenomenon showed that when infants were habituated to a hand reaching towards an object, they were subsequently surprised when the hand continued reaching in the same trajectory if the object was no longer present; instead they expected the hand to continue reaching towards that object even when its location had changed, indicating that infants conceived of the hand’s action as being intentionally directed towards the object (Woodward, 1998). This basic effect has been replicated many times (Leslie & Biro, 2007; for a review see Woodward, Sommerville, Gerson, Henderson, & Buersh, 2009) and extended to actions such as looking (Phillips, Wellman, & Spelke, 2002), pointing (Behne, Carpenter, & Tomasello, 2005; Liebal, Behne, Carpenter, & Tomasello, 2009; Woodward & Guajardo, 2002), reaching at a distance (Brandon & Wellman, 2005), and tool use (Sommerville, Hildebrand, & Crane, 2008), all of which infants seem to interpret as goal-directed by the first or second year of life.

Infants also seem to make social evaluations at a very young age. Infants can distinguish between positive social acts and negative ones (Premack & Premack, 1997) and expect an agent to prefer those that help it over those that hinder it (Kulmeier, Wynn & Bloom, 2003).
Do infants use their ability to infer the intentions of others when they make social evaluations? Or do infants rely on more low-level, behavioral cues to make social evaluations, implying that the capacity to use intention information may not emerge until much later in life? A growing body of work suggests that infants do in fact evaluate intention in the process of social reasoning.

One of the most foundational works in this series was conducted by Kuhlmeier, Wynn, and Bloom (2003). Infants (12-months and 5-months) were habituated to videos showing a circle attempting to climb a hill. In separate scenes, the circle is helped up the hill by a square and pushed down the hill by a triangle. At test, infants saw the circle approach the helping shape or the hindering shape. Twelve-month old infants preferred to look at the event in which the circle approaches the helper. Five-month old infants showed no preference between the test events. This study suggests that by twelve-months, infants expect that agents will develop preferences for those who are pro-social (those who provide assistance to the achievement of a goal) and will develop aversion to those who are anti-social (those who encumber a goal). A rich interpretation of these results suggests that second-order theory of mind is at play. That is, the circle has a goal – climbing the hill – and the other shape has a goal concerning the circle’s goal – to help or hinder the circle in its hill-climbing pursuit.

Follow-up work in this paradigm suggested that infants themselves develop preferences for agents after witnessing third-party interactions (Hamlet et al., 2007). Infants (6- and 10-month olds) saw a puppet show in which a circle attempted to climb a hill. In one version of the show, a hindering shape pushed the circle down the hill. In the other version, a helping shape pushed the circle up the hill. Infants were then presented
with the wooden shapes from the puppet show and asked “Would you like to pick a toy?” Infants preferred the helping shape over the hindering shape. Furthermore, when cues to agency were removed from the climber so that the climber was seen as an inanimate ball, infants did not exhibit a preference for the “helping shape” (the shape that pushed the ball up the hill) over the “hindering shape” (the shape that pushed the ball down the hill). That is, when the shape was not perceived as having a desire, then no social attribution was made with regard to the other shapes and infants chose randomly between them. A similar finding was observed with 3-month-old infants using a preferential looking measure (Hamlin, Wynn, & Bloom 2010).

Do infants indiscriminately prefer prosocial to antisocial others, or are they able to integrate a representation of the character of the moral patient into their evaluation of the appropriate way to treat that agent? Put another way, do infants understand that it is sometimes preferable to harm a harmer rather than help a harmer? Hamlin, Wynn, and Bloom (2011) investigated this question. 5-month and 8-month-old infants were shown a puppet show that involved a helper facilitating an agent’s goal and a hinderer obstructing an agent’s goal. Infants were then shown a third character interacting with the helper or the hinderer. This third character either took a ball away or gave a ball to the original helper/hinderer. Infants are then given the opportunity to choose to play with the Giver or the Taker. At 8 months of age, when infants saw the Giver and Taker interacting with the Helper, infants preferred the Giver over the Taker. However, when infants saw the Giver and Taker interacting with the Hinderer, infants preferred the Taker. That is, infants preferred those who punish antisocial others and reward prosocial others. In contrast, at 5-months of age, infants preferred the Giver in both conditions, indicating a
difficulty integrating the identity or previous behavior of the moral actor into their representation.

The studies reviewed above indeed seem to suggest that infants are using information about the intentions of agents to make social evaluations. However, deflationary accounts have been offered of these results, suggesting that infants could succeed on even very complex tasks using behavioral cues alone and not mental state reasoning (eg: Ruffman & Perner, 2005).

In response to these concerns, Hamlin and colleagues (2013) developed a study in which subjects’ judgments could not be accounted for by behavioral cues because the conditions (helper vs hinderer) involved the exact same actions and only differed on mental states of the agents. In addition, the authors developed four computational models based on modified Bayesian inverse planning (Baker, Saxe & Tenebaum, 2009) to test against the data. Each model assumes that infants use more mental state information when evaluating the actions of the agents. The models make different predictions for each condition, allowing the experimental data to differentiate between them. The models include (1) a full mental model, which assumes that infants understand second-order goal attribution, (2) a mid-level model, which assumes that infants use second-order goals but not second-order beliefs, (3) a goal-completion model, which uses goal information but does not correctly use preferences to infer goals, and (4) feature-based model which does not rely on mental states. Results show that the full mental model explain the data better than any of the other three models. This is strong evidence that infants use intention information in making social evaluations.
3. Conclusion

I will conclude this review by pointing out several questions that arose as posing particular challenges to the field of moral cognition.

- How are intentions inferred for novel actions in the moral domain?
- There seems to be strong evidence that infants use intention to make social evaluations? Does this vary cross-culturally? If not, how/why does the reliance on intention information in adults vary so drastically across cultures in the domain of moral judgment?
- Do infants make moral judgments, or just social evaluations? What methods would enable us to answer this question? What counts as a moral judgment anyway?
- If infants can use intention to make social evaluations, then why do young preschoolers (3-4 year olds) seem to rely on outcome and ignore intention when making moral judgments?
II. Presumed Innocent? How Tacit Assumptions of Intentional Structure Shape Moral Judgment

1. Introduction

Agents move through the world constantly starting causal sequences of events that can be parsed in infinite ways. For example, I raise my hand signaling that I want to answer a question, but at the same time catch the eye of a student sitting behind me, whack someone who suddenly got up from her chair, create a small breeze, and increase the number of hands raised by one. Most on-lookers would immediately infer that the former of these effects was the one that I intended, yet, the bare evidence supports inferences that any of them may have been the agent’s goal. How do we, from a very young age, reliably determine the goal of agents, given the vast number of effects every action causes?

Some of the effects caused by an action seem to immediately leap out as good candidates for being a goal (e.g. obtaining an object, arriving at a location, communicating, harming or helping someone). We will refer to such effects as “salient” effects, though we put aside the question of how salience gets attached to certain effects and not others. We instead focus on the problem of how we infer the intended goal of an action when more than one salient effect occurs.

Determining which (salient) effects an agent intends is central to our capacity to make moral judgments. In cases where some of the effects of an action are morally good and some are morally bad, our moral judgments can change dramatically depending on whether the morally good or bad effects were intended (Young, Cushman, Hauser & Saxe, 2007; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; Hamlin, Ullman, Tenenbaum, Goodman & Baker, 2013). For example, if I raised my hand in order to
whack the person who suddenly got up from her chair, I may be judged more morally culpable for the harm I caused her than if my whacking of her was accidental and my intended goal was to signal my eagerness to answer a question.

We review several theories of the development of intention inference, pointing out how each falls short of explaining how we can infer the intention of novel actions with more than one salient effect. We suggest a solution to this problem that is particular to the moral domain, a good intention prior. Finally, we present the results of two studies – one with adults and one with preschoolers – that suggest that subjects solve the goal inference problem for novel actions in the moral domain using the good intention prior.

1.1 Goal inference theories

First, we should differentiate between two kinds of intention (Premack, 1990; see also Searle’s distinction between intention-in-action and prior intention, 1983). First, someone can act intentionally when he acts deliberately or non-accidentally. Simply observing a white dot on a black background change direction and speed is perceived as intentional action done by an animate agent (Temoulet & Feldman, 2000; see Scholl & Tremoulet, 2000 for a review), though perceiving this sort of intentionality does not necessarily require representing that the agent is acting towards a particular goal (but see Pantelis & Feldman, 2012). Second, an intention can refer to a plan of action that has a causal connection to the behavior of an agent, predicated on certain beliefs, and aimed at bringing about a certain goal (e.g., Bratman, 1987). Our focus is on theories that describe how we infer the latter sort of intention, in particular how we infer goals given the observation of an action.
There is now a significant body of evidence that by 6 months old, infants are capable of viewing the actions of agents as goal-directed (Woodward, 2013). Early studies of this phenomenon showed that when infants were habituated to a hand grasping one of two objects, they were surprised when the hand reached for the other object but not when the hand followed a new path to the first object (Woodward, 1998). This basic effect has been replicated many times (Biro & Leslie, 2007; for a review see Woodward, Sommerville, Gerson, Henderson, & Buresh, 2009), though it remains an open question how we develop the capacity to see actions as goal-directed.

Woodward (2013) suggests that an action may be perceived as goal-directed if the infant herself has intentionally performed that action with a particular goal in mind. Infants also understand that individual actions can be related to each other based on their role in bringing about some overarching goal (e.g., Woodward & Sommerville, 2000). That is, if an infant observes a sequence of novel actions that culminates in a familiar goal (such as obtaining an object), and the novel actions are causally connected to the goal (based on the psychical and psychological constraints of the action context), then the novel actions can be perceived as intentional means to the familiar end (Woodward & Sommerville, 2000).

This is a plausible suggestion for how we infer the goal of a novel action in cases where one salient effect is caused by the action. However, it is less clear how infants (or adults) would be able to interpret a novel action that could be perceived as being part of two different causal and intentional sequences. For example, if an action simultaneously results in obtaining an object and making a fun sound, did the actor intend to obtain the object, create the fun sound, or both (cf. Sommerville & Woodward, 2005)?
Gergely and Csibra argue that infants in the first year of life interpret the actions of agents by applying a teleological relationship between three aspects of reality: actions, goal states, and situational constraints (Csibra & Gergely, 1998; Gergely & Csibra, 2003; Csibra, Bíró, Koós, & Gergely, 2003; see also Scott & Baillargeon, 2013). The principle of rational action relates these into a teleological schema by assuming that actions realize goal-states in the most efficient way possible. For example, infants were habituated to an agent jumping over a wall and ending up at a goal object. At test, the wall is removed. Infants look longer (indicating their surprise) when the agent again follows the curved path to the goal object as compared to a case where the agent follows a straight path to the goal object (Gergely, Nádasdy, Csibra & Biro, 1995). This suggests that infants interpret the action of the agent in the habituation phase as the most efficient action available to reach the goal state. At test, when the wall is no longer present, infants expect the agent to take a new path given the new environmental constraints, but to continue to act in a goal-directed, efficient manner.

On this account, the goal of a novel action is inferred by determining whether it is an efficient means to bring about any of the effects that have been observed. However, where more than one effect is brought about in the most efficient manner, Csibra and Gergely point out (2007), additional cognitive constraints are needed to determine the goal of the action.

Biro & Leslie (2007; see also Leslie 1991, 1994) propose that infants are innately equipped with a capacity to pick out goal-directed action based on certain motion cues, including equifinality, action-effect pattern, and especially self-propelled motion (DiGiorgio et al., 2016). A domain-specific learning mechanism can then detect statistical
regularities about the surface-level features of the objects that typically exhibit the cues, e.g. hands. Once agents are identified, infants can learn to infer their goals by keeping track of what those effects those agents typically cause. Critically, this mechanism does not address the critical question of how infants (or adults) might determine the goal of a novel action they observe for the first time.

Meltzoff (2005; 2007) suggests that infants develop an understanding of other minds through a “like-me” comparison. When infants see others acting in ways that the infants have acted in the past, infants recognize that the other is “like-me” and can project the mental state that went along with the action onto the agent they are now observing. The “like-me” framework has difficulty explaining how children determine the goals of actions that they have never performed, do not have the motor skills to carry out, or are not performed by conspecifics (e.g. in cases of shapes moving on a screen; Csibra & Gergely, 1998).

Baron-Cohen (1997) (extending Premack, 1990) suggests that early in infancy we may be equipped to recognize certain pre-defined goals such as freedom, companionship, arousal, arriving at a certain endpoint, effecting another agent, and reciprocating from a previous interaction (Premack, 1990). This theory does not have a way of dealing with novel actions that fall outside the schemas infants are prepared to deal with.

Above we reviewed theories of how infants before the first year of life infer intention. How does this relate to the capacity of older children and adults to infer intention? Some theories seem to suggest that the core mechanism that infants use to infer intention maintains its status as the central mechanism of interest through adulthood. For instance, Baker and colleagues (Baker, Saxe, & Tenenbaum, 2009; Baker, Tenenbaum, & Saxe,
2006; Pantelis & Feldman, 2012) built and tested formal computational models of the teleological stance (Csibra & Gergely, 2007), finding evidence that this mechanism still describes the adult ability to infer goals. On the other hand, some theories explicitly deny that the infant theory of intention is at all continuous with the adult theory (Premack, 1990). Other theories suggest that some core information is present very early on, which enables a learning sequence to take place, vastly increasing the ability to infer intention through late infancy and possibly into childhood and beyond (Biro & Leslie, 2007; Meltzoff, 2007; Woodward, 2013).

Each theory that attempts to explain how the adult intention inference system could develop must be able to address the issue of how we can infer the goal of novel actions. Our proposal is that domain-specific knowledge may be critical in enabling observers to infer the goal of a novel action with multiple salient effects. This knowledge could take the form of prior distributions over the effects that are more likely to be intended. We may be able to learn which effects are more likely to be intended by building up priors in contexts where actions result in just one salient effect (e.g. a simple act of helping or hindering).¹

There is already some evidence to suggest that domain-specific priors may help infants disambiguate the goals of novel actions with multiple salient effects. Sommerville & Crane (2009) presented 10-month old infants with an action sequence that is ambiguous for infants of this age. Infants saw an experimenter pull a cloth to bring a toy that was resting on it into reach. Previous work has shown that 10-month olds

¹ We use the term “prior” loosely throughout this section to mean something like an assumption, default, or bias. A more careful treatment of the nature of this bias would present calculations to show how a Baysian prior would function in a probabilistic application, integrated with other elements of the relevant cognitive processes. That treatment is beyond the scope of the present work.
can interpret this action in at least two distinct ways, as directed towards the obtaining the toy or as directed towards the cloth (Sommerville & Woodward, 2005). The insight of Somerville & Crane (2009) is that 10-month olds could be encouraged to interpret the cloth-pulling behavior as being directed towards the toy if they were previously shown the experimenter reaching for and obtaining the toy in a non-ambiguous context. These data make the highly intriguing suggestion that infants can use prior knowledge about the goals and/or preferences of a particular agent to disambiguate the goal of a novel action for that agent.

By 12-months of age, infants no longer find the cloth-pulling action to be ambiguous and they interpret that action as being directed towards obtaining the toy (Sommerville & Woodward, 2005). What has happened between 10- and 12-months of age, such that 12-month olds no longer need to be given a prior expectation about the goals of a particular agent to infer that agent’s goal? It seems plausible that 12-month olds already have this prior expectation and that the prior can now be applied to any agent. Put in our terms, 12-month olds have acquired domain specific knowledge about the goals of agents with respect to objects, allowing them to solve the inference problem for novel actions with multiple salient effects in this domain by drawing on their prior.

In the next section, we suggest a solution for the problem of goal inference for a novel action with multiple salient effects in the moral domain. Our suggestion is this: when there are two morally charged effects, one good and one bad, the prior probability favors the inference that the agent intended the good effect and not the bad effect (Mikhail, 2007). Put another way, a unique solution to the goal inference problem for
novel actions in the moral domain can be achieved by positing particular domain-specific knowledge: a *good intention prior*.

1.2 Goal inference in moral cognition

Interpreting the goal of novel actions is critical to making moral judgments about those actions. For actions that have both morally good and morally bad effects, the agent’s intention concerning those effects may be critical in determining the moral permissibility of the action. Many theories of moral cognition highlight the role that intention plays in making moral evaluations for adults (Young, Cushman, Hauser & Saxe, 2007; Mikhail, 2011; Greene, 2013; Cushman, 2013; Malle, Guglielmo, & Monroe, 2014; for a review see Doris, 2010) as well as children (Cushman, Sheketoff, Wharton, & Carey, 2013; Baird & Astington, 2004; Killen, Mulvey, Richardson, Jampol & Woodward, 2011; for a review see Killen & Smetana, 2015), and potentially infants (for a review, see Hamlin, 2015). Yet the question of how we attribute an intention to an agent in morally-charged cases where multiple intention ascriptions are possible has gone largely unremarked upon (as noted by Mikhail, 2007).

A significant percentage of research in moral psychology has involved “trolley problems”, which bring out the extent of this problem (Mikhail, 2002; Cushman, Young & Hauser, 2006; Pellizzoni, Siegal & Surian, 2010; Waldmann & Dieterich, 2007; Schwitzgebel & Cushman, 2012; Greene, Cushman, Stewart, Lowenberg, Nystrom, & Cohen, 2009; for a review see Waldmann, Nagel & Wiegmann, 2012). In a classic trolley-problem case, a train has gone out of control and threatens the lives of innocent people stranded on the tracks. An agent intervenes, frequently by redirecting the train,
causing the originally threatened individuals to be saved and different people to be killed. The act of redirecting the train arguably counts as novel action which leads to two salient moral effects: some people are saved while others are killed. How do subjects, who are typically only given information about the causal sequence of events that occur, infer the intention of the agent? Were the deaths intended? The lives saved? Both? (See Figure 1.) This question is particularly stark in the trolley-problem scenario, but is equally pressing for cases of moral action that involve multiple morally charged effects.

Figure 2.1: When one action has two effects (for instance, one good effect and one bad effect), a single causal structure is compatible with two different intentional structures. In this figure, the red line indicates the agent’s action plan – the sequence of actions she
intends that will bring about her goal. These diagrams should be read from bottom to top: the agent first brings about a basic [Act-token] and then two effects result: one good and one bad. As shown here, it is possible that the agent intends the good effect and that the bad effect is a foreseen but unintended side-effect of the basic act-token (possible intention structure #1). The reverse possibility is also a viable transformation of the causal structure (possible intention structure #2).²

Because the theories of goal inference fall short of being able to explain how we infer the goal of a novel action with multiple salient effects, theories of moral cognition cannot simply hope to “plug in” a theory of goal inference as a solution to this problem for moral judgment. We propose that there is a solution to the problem of goal inference for a novel action that is particular to the moral domain and has explicit moral content.

While some theories of moral cognition that emphasize the role of intention in moral judgment have ignored the problem of intention inference entirely (often by explicitly telling subjects what an agent intends; e.g. Young & Saxe, 2011; Cushman, Sheketoff, Wharton & Carey, 2013; Baird & Astington, 2004), others have attempted to describe the way that action representations are built from impoverished stimuli (i.e. lacking explicit intention information) so that moral judgment can proceed. The theories in this latter category, while not always acknowledging the difficult nature of the problem of intention inference for novel actions, seem to have either tacitly or explicitly assumed something like a good intention prior – though this assumption lacks empirical support.

² The “act trees” used in this and subsequent figures are borrowed from notation for describing action used by Mikhail (2011, 2007), who in turn borrowed them from Goldman (1970). These diagrams are loose hypotheses about the content and structure of our mental representations of action. The cognitive science of action representation is just beginning to emerge (e.g. Shipley & Zacks, 2008), but knowledge gained from studies in that field would greatly inform how these diagrams should be built to more rigorously reflect the mental processes at work in representing action.
The Moral Grammar Hypothesis (Mikhail, 2007; 2011) suggests that deontic judgments are the output of modular systems that runs a series of computations over highly structured, informationally rich mental representations. Mikhail (2007; 2011) notes that constructing an intentional structure given only the input available from the prompts he provides subjects requires making an assumption about which of the morally good or bad effects in the causal sequence of events count as the goal of the actor. Mikhail’s suggestion is that we do this by way of the good intention prior.

While no other theory has made this explicit, several seem to tacitly assume the good intention prior. For example, Greene (2013) suggests that we have a modular system that inspects the action plans of agents as part of the mechanism of moral judgment. Despite his emphasis on the importance of processing the action plans of agents, Greene does not provide an account of how we infer which effects count as side-effects, means, and goals. However, in each of the action plans that he uses to illustrate how we represent the mind of a moral agent in a particular dilemma, he indicates that saving lives is the goal of the agent. In the background of Greene’s theory, therefore, is the critical, untested assumption that we assume that agents intend the good effects and not the bad effects.

Two recent theories (Cushman, 2013; Crockett, 2013) suggest that mechanisms of moral judgment can be described using a dual-process approach that is instantiated by “model-free” and a “model-based” reinforcement learning systems, which can be mapped onto action- and out-come based value representations, respectively. Drawing on recent developments in hierarchical reinforcement learning models, these theories show how the two systems can be integrated to allow intention to be represented as hierarchical
superordinate/subordinate-goal planning. Value can then be applied to the representations to yield moral judgments in line with, for example, the means-side-effect distinction. However, in order for this process to even get off the ground, an assumption exists in the background that the agent acts “out of concern for others rather than malice” (Cushman, 2013, pg. 283). Given this assumption, goals and sub-goals can be assigned to the agent and moral cognition can proceed. Again, that that assumption is empirically warranted remains an open question.

2. Experiment 1

We suggest that in trolley-like scenarios in which no intention information is presented, subjects use the Good Intention Prior to select one intention structure and reject the other. Therefore, if intentions are added to the stimulus that explicitly indicate which intention structure to choose, we expect there to be no difference in how subjects judge the case if the intentions provided match those assumed by the prior. However, if the intentions provided are different from those assumed by the prior, subjects' judgments may change to reflect the new intention information.

2.1. Methods

Subjects read a story in which a train is about to kill five people who are standing in its path. In response, an agent throws a switch, thereby preventing the train from killing the five people and with the same action causing the train to turn down a side-track and kill one person. (For text of the stimuli, see Appendix B.) The causal structure of the agent’s action was unambiguous. However, two intention structures are compatible with the causal structure: it is possible that the agent’s intention was to save the five people (and that the harm to the one person was a foreseen but unintended side-effect) or that the
agent intended to harm the one person (and that saving the five was a foreseen but unintended side-effect). (Figure 2)

Figure 2.2. The relationship between the causal structure and the intention structures of the story in Experiment 1. The causal structure of the morally-charged story is compatible with two intention structures. In the Bad Condition and Good Condition, information is provided which allows subjects to choose one of the intention structures. In the Neutral Condition, no intention information is explicitly given.

Subjects received the story in one of three conditions. Subjects in the Neutral Condition received no explicit information about the agent’s intention. Subjects in the Good Condition received information that the agent intended the good effect of his action (saving the five people on the main track). Subjects in the Bad Condition received
information that the agent intended the bad effect of his action (killing the one person on
the side-track). Subjects were then asked two test questions. First, they were asked to
issue a deontic judgment of the agent’s action: “Is it morally permissible for Hank to
throw the switch?” Second, they were asked to judge the agent’s intention by responding
“true” or “false” to the following statement: “Hank threw the switch in order to kill the
man.” (Subjects also responded to a series of additional true/false questions about the
sequence of events in the story narrative, which are not relevant to the present analysis.)

2.2. Subjects

We stopped data collection when one hundred sixty-six subjects (recruited from
Amazon’s Mechanical Turk) completed the study in order to achieve sample sizes
comparable to previous studies that have used similar methodology (cf. Greene et al,
2009). Thirty subjects were excluded from analysis for failing to follow the instructions.
All subjects were paid for participating.

2.3. Statistical Analyses.³

We assume that imputed intention is an important determinant of moral judgments,
with the consequence that subjects’ judgments will be negatively affected by the
imputation of bad rather than good intentions in the agent. Our hypothesis is that in the
Neutral Condition, when intentionality is not specified, intention to bring about the good
effect is assumed, with the consequence that moral judgments will be the same in this
condition as in the condition where good intentions are stipulated. Conventional statistics
are not suited to the assessment of this second hypothesis, because it is a null hypothesis.

³ We are grateful to Randy Gallistel for his expert help writing this section and with computing Bayes
Factors.
In the conventional formulation of statistical inference, the failure of the null hypothesis ($H_0$) to predict the data well is taken to license the conclusion that the data support an alternative hypothesis ($H_1$), but $H_0$ is not quantitatively specified and so *a fortiori* not tested against the data. In this formulation data can never be taken to support a null hypothesis. In the Bayesian formulation of the inference problem, there are (at least) two quantitatively formulated hypotheses. One computes the relative likelihood of the competing hypotheses given the data (the Bayes Factor). We computed both Bayes Factors and conventional *p* values.

In computing Bayes Factors, we considered two alternatives to our null hypothesis: The first is that the probability of a given moral judgment in the Good condition provides no information about judgments in the Neutral condition. On this alternative, the probability of a judgment in the Neutral Condition may with equal probability assume any value within the obtainable range. This is the simplest formulation of what the implicit alternative to the null is when one does a 2-tailed *t*-test for difference in the means. A more refined alternative is that the judgments and ratings in the Neutral Condition will be more negative than in the Good Condition, because bad intentions are imputed to the agent by a few or even all the subjects. On this alternative, the probability of a favorable deontic judgment or of a good/bad rating in the Neutral Condition may with equal probability assume any value on the negative side of the value in the Good Condition. This is the simplest formulation of what the implicit alternative is in a 1-tailed *t*-test.

A more or less conventional interpretation of the support a Bayes Factor of a given magnitude provides for the favored (odds on) hypothesis is: $< 2 = \text{trivial support}; 2 \text{ to } 3 =$
weak support; 3 to 10 = moderate support; 10 to 100 strong support; >100 = decisive support. This support is always relative to the specified alternative; when the Bayes Factor in favor of the alternative to the null is 100, then the odds are 100:1 that the alternative is better than the null—and vice versa! With modest sample sizes and a plausibly restricted alternative hypothesis, it is impossible to obtain really large Bayes Factors in favor of the null even when the null predicts the data perfectly; whereas when the null predicts the data badly, factors in the millions may be obtained for alternatives to it, alternatives that predict the data better.

2.4. Results

First we will consider subjects’ deontic judgments. In the Neutral Condition, 78% of subjects (35 out of 45) judged the case permissible. In the Good Condition, 78% of subjects (35 out of 45) judged the case permissible. In the Bad Condition, 46% of subjects (21 out of 46) judged the case permissible.

There was no significant difference between the Neutral and Good Conditions. (The number of subjects judging the cases as permissible is identical in the Neutral and Good Conditions, so classical statistical tests are unnecessary to support this claim.) The Bayes Factors favored the null (6.82 1-tailed; 4.64 2-tailed). In contrast, there was a significant difference between the Neutral Condition and the Bad Condition Upton’s $\chi^2 (1, n = 91) = 9.81, \varphi = 0.33, p=.0017$, two-tailed, with Bayes Factors decisively in favor of the alternative to the null whether it was 1- or 2-tailed. (See Figure 3.)
Next, we will consider subjects’ intentionality judgments. In the Neutral Condition, 82% of subjects (37 out of 45) judged that the agent did not intend the bad effect of his action. Likewise, in the Good Condition, 82% of subjects (37 out of 45) judged that the agent did not intend the bad effects of his action. In the Bad Condition, 35% of subjects (16 out of 46) judged that the agent did not intend the bad effect of his action.

There was no significant difference between the Neutral and Good Conditions. In addition, the Bayes Factors favored the null (7.82 1-tailed, 5.03 2-tailed). In contrast, there was a significant difference between the Neutral and Bad Conditions, Upton’s $\chi^2$ $1, n = 91$ = 20.82, $\varphi = .48$, $p < .001$, two-tailed, with Bayes Factors decisively in favor of the alternatives to the null, whether 1- or 2-tailed. (See Figure 4.)
Finally, Pearson Correlation analysis was conducted on the two dependent variables, revealing that deontic judgment and intentionality assessment were significantly related, $r = -.26$, $n = 136$, $p = .003$, two-tailed.

3. **Experiment 2**

Does the good intention prior take years of social learning to emerge in adults? Or is it present much earlier in development? Experiment 2 was designed to test whether, preschoolers, like adults, use the Good Intention Prior when confronted with trolley-like cases.

3.1. **Methods**

Subjects were tested individually in quiet locations in their preschools or in the lab. Subjects were first trained on a Likert scale, with X’s on one end and stars at the other. Children were taught that the ends of the scale could be used to talk about things that
were “really bad” and “really good” and that the intermediate points were for things that were “a little bad” and “a little good”, with the point in the middle being for things that were “just OK.” Children were guided in practicing with the scale. Then, children were told stories in which a simple morally good or bad action took place. Children were asked to issue a moral judgment of the action (“Should he/she have done that?”) and were asked to rate the action on the Likert Scale. Only children who expressed competence making simple deontic assessments and using the scale to describe moral behavior were tested further (see Appendix A for further details).

Note that our “should” question is simply a proxy for measuring deontic judgment in children, who usually are not yet competent with the terms “morally permissible” and “morally impermissible”. While, our data suggest that children respond to the question “should she have done that” in a similar way that adults respond to straight-forward questions of deontic judgment, it remains an open question whether “should” captures a slightly different concept in children than “morally permissible” captures in adults.

Children were then told a story similar in structure to the adult story. In the story, a girl prevents a squirrel from eating five children’s cookies (by putting up a gate) and with the same action causes the squirrel to eat one child’s cookie. Just like in Experiment 1, the causal structure of the girl’s action was unambiguous. However, two intention structures are compatible with the causal structure: it is possible that the girl’s intention was to save the five children’s cookies (and that the harm to the one child was a foreseen but unintended side-effect) or that the girl intended to cause the squirrel to eat the one child’s cookie (and that saving the five was a foreseen but unintended side-effect).
Subjects received the story in one of three conditions. Subjects in the Neutral Condition received no explicit information about the agent’s intention. Subjects in the Good Condition received information that the agent intended the good effects of her action (saving the five children’s cookies). Subjects in the Bad Condition received information that the agent intended the bad effects of her action (causing the squirrel to eat the one child’s cookie). A series of control questions were asked to ensure subject memory and comprehension of the story (see Appendix B). Children were then asked three test questions. First, they were asked to judge the agent’s intention: “Did Sally make this one kid sad on purpose?” Second, they were asked to issue a deontic judgment of the agent’s action: “In this story Sally used her gate. Should she have done that?” Finally, subjects were asked to rate the actor’s action on the Likert Scale: “Was what Sally did good, bad, or just OK?” (See Appendix B for text of stimuli.)

3.2. Subjects.

Fifty children between the ages of 37 months and 72 months received the Neutral Condition (M = 55.6 mo; SD = 9.0 mo), 29 of which were girls. Forty-seven children between the ages of 40 months and 72 months received the Good Condition (M = 53.8 mo; SD = 8.2 mo), 23 of which were girls. Thirty-six children between the ages of 42 months and 68 months received the Bad Condition (M = 56.3; SD = 8.1), 17 of which were girls. 37 additional children were excluded from the study, 32 for failing scale training, 2 for failing to cooperate, 1 for failing control questions, 1 for parent interference, and 1 for experimenter error. We stopped data collection after testing 170 children in order to reach sample sizes that have been previous used for studies with a similar methodology (cf. Saunders, 2014).
3.3. Results

First we will consider subjects’ deontic judgments. In the Neutral Condition, 60% of subjects (30 out of 50 subjects) judged the case permissible – that is, they responded “yes” to the question “Should she have done that?” In the Good Condition, 72% of subjects (34 out of 47) judged the case permissible. In the Bad Condition, 22% of subjects (8 out of 36) judged the case permissible.

There was no significant difference between subjects’ responses to the Neutral and Good Conditions, Upton’s $\chi^2(1, n = 97)= 1.62$, $\varphi =.13$, $p = .202$, two-tailed. In contrast, there was a significant difference between the Neutral and Bad Conditions, Upton’s $\chi^2 (1, n = 86) = 11.96$, $\varphi =.37$, $p <.001$, two-tailed. The 1-tailed BF for the Good-Neutral comparison was 1.47 in favor of the null; the 2-tailed BF was 1.90 in favor of the null. The BFs for the Good-Bad and Neutral-Bad comparisons were all decisive (>100) in favor of the alternative hypothesis. (See Figure 5.)

![Figure 2.5](image)

Figure 2.5: Preschool subjects’ moral judgment of the action of the agent. Subjects answered the question, “Should she have done that?”
Next, we will consider subjects’ Likert ratings of the action of the agent. Likert scale ratings were scored as follows: really bad = -2; a little bad = -1; just OK = 0; a little good = 1; really good = 2. Subjects in the Neutral Condition and the Good Condition both rated the agent’s action as slightly above the midpoint of the scale (Neutral: $M = .24$, $SD = 1.36$; Good: $M = .23$, $SD = 1.31$). Subjects in the Bad Condition rated the agent’s action as bad ($M = -1.11$, $SD = .98$). Analysis of variance revealed a significant difference between the conditions, $F(2, 130) = 15.27$, $\eta^2 = .19$, $p < .001$. Planned pairwise comparisons revealed that there was no significant difference between the Neutral and Good Conditions (independent-sample $t$-test, two-tailed, $t(95) = .022$, $r = .0022$, $p > .250$). Furthermore, the 2-tailed BF for the Good-Neutral comparison was 5.65 in favor of the null. In contrast, there was a significant difference between the Neutral Condition and the Bad Condition (independent-sample $t$-test, two-tailed, $t(84) = 5.24$, $r = .50$, $p < .001$). The BFs for the Good-Bad and Neutral-Bad comparisons were all decisive ($>100$). (See Figure 6.)

![Figure 2.6: Preschool subjects’ Likert ratings of the action of the agent (error bars are 95% confidence intervals).](image)
Next, we will consider subject’s intentionality judgments. In the Neutral Condition, 66% of subjects (33 out of 50) judged that the agent did not intend the bad effect of her action. In the Good Condition, 60% of subjects (28 out of 47) judged that the agent did not intend the bad effect of her action. In the Bad Condition, 14% of subjects (5 out of 36) judged that the agent did not intend the bad effect of her action.

There was no significant difference between the Neutral and Good Conditions, Upton’s $\chi^2 (1, n = 97) = .42, \varphi=.07, p>.250$, two-tailed, and the BFs show that the results favor the null hypothesis (7.20 1-tailed, 3.37 2-tailed). In addition, there was a significant difference between the Neutral and Bad Conditions, Upton’s $\chi^2 (1, n = 86) = 22.77, \varphi=.51, p<.001$, two-tailed and the BFs decisively favored the alternative hypothesis (>100) whether it was 1- or 2-tailed. (See Figure 7.)

Figure 2.7: Preschool subjects’ judgments of the intention of the agent. Subjects answered the question, “Did she make the one kid sad on purpose?”

Finally, correlation analysis was conducted on the three dependent variables (should judgment, Likert rating, and intentionality assessment) for $n = 133$ subjects. As
anticipated, the dependent variables are significantly correlated. The correlations between each pair of variables are reported in Table 1.

<table>
<thead>
<tr>
<th>Should</th>
<th>Action Rating</th>
<th>Intention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Should</td>
<td>--</td>
<td>0.52**</td>
</tr>
<tr>
<td>Action Rating</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Table 2.1: Correlation matrix for should, Likert rating, and intention questions. Pearson correlations are reported. \( n = 133; * = p=0.011, \) two-tailed; \( ** = p<0.001, \) two-tailed.

4. Discussion

Theories of intention inference have not been able to explain how we can infer the goal of a novel action when multiple salient effects are observed. We suggest that domain-specific priors over goals (developed in less ambiguous cases) can help solve the problem. If this is the case, then it is an empirical question of each specific domain what goals are favored by the prior probability. Priors in some domains may vary dramatically based on individual experience, while some are likely to be more consistent across individuals and groups.

The problem of how to infer the intention of a novel action is particularly important in the moral domain. When a novel action results in both morally good and morally bad effects, determining which of them the agent intends is critical for making a moral judgment. Theories of moral cognition, even those that highlight the important role that intention plays in moral judgment, have either (1) ignored this issue and begun their theoretical work given that the subject has already inferred intention (e.g. Young & Saxe, 2011; Cushman et al., 2013) or (2) made an explicit or tacit assumption about subjects’
intention inferences that has not been empirically validated (Mikhail, 2011; Greene, 2013; Cushman, 2013; Crockett, 2013). Given that we think that the problem of inferring the intention of novel actions is an empirical question for each domain of cognition, we think that theories of moral cognition should not simply ignore this question (as in strategy (1)) or make empirically invalidated claims (as in strategy (2)).

We propose an empirically testable solution to the problem of inferring intention for novel action in the moral domain: a good intention prior, namely, when a novel action is observed that results in morally good and bad effects, the prior probability favors the good effect as the actor’s goal. If this hypothesis is correct, then in our studies, when no intention information is available (Neutral Condition), subjects should treat the morally good effect as the goal of the agent’s action (as they do in the Good Condition). If this hypothesis is incorrect, subjects may (1) be more likely to treat the morally bad effect as the goal (as compared to the Good Condition) or (2) choose equally between the two options.

The principle finding of our studies is this: when no intention information is explicitly stated for a novel action (Neutral Condition), adult and preschool subjects judge the case in the same way they judge cases in which subjects are explicitly told that the agent intends the good effect of the action (Good Condition). This suggests that in our Neutral Condition (and the vast majority of trolley-like tasks) subjects are deploying a Good Intention Prior, supplying missing intention information by assuming that the agent intends the good effects and not the bad effects of her action. In contrast, when the story contained information that the agent intended the bad effect, there were significant
differences on measures of intention and deontic status as compared with each of the other two cases (Neutral and Good).

Our cases were designed to approximate the state of having no information about an agent’s intention. Learning additional information about a particular agent, his relationship to the other agents in the scenario, his history and motives could all impact the prior, ultimately leading an observer to conclude bad or ambiguous intentions. Moreover, the subjects we tested generally came from relatively stable environments (preschoolers were mostly residents of Middlesex County, NJ). Being exposed to more intentional harm on a regular basis would have the potential to push around individuals’ priors on the intentions of agents in general.

We have already drawn attention to the fact that trolley-like cases are a prime example of a novel moral action that has both good and bad effects. We have pointed out that findings from these sorts of cases (eg, that the “side-track” case is judged morally permissible) seem to suggest that subjects are using a good intention prior in their moral evaluations (to disambiguate intention structures). Here, we point to one additional study which also seems to suggest, using a different paradigm, that subjects (here, children) drawn on the good intention prior to make moral judgments. Killen and colleagues (2011), told subjects (3-8 years old) that a child put a cupcake in a bag and then went outside to play. Later, another child is helping a teacher clean up the room and throws the bag in the trash. This story involves two morally-charged effects: throwing away another child’s cupcake (a bad effect) and helping the teacher clean up the room (a good effect). Children are asked to assess the intention of the protagonist and make a moral judgment about his action. No explicit intention information is given in the story;
children are required to infer the intention of the actor. The central finding is that when children gain false-belief understanding (measured by ability to pass a verbal false-belief task), they are less likely to attribute negative intentions to the transgressor (indicating that the intention of the actor was to bring about the good effect). We suggest that children attribute a false belief to the actor in this story as the best possible explanation of the agent’s behavior given their prior over good intentions. Without this prior, an entirely sensible read of the story is that the actor knowingly and maliciously discarded his classmate’s cupcake.

Two sets of recent findings may seem, on the face of things, to contradict our proposal for a good intention prior. First, it has been suggested that there may be a “negativity bias” in agency attribution: infants and adults seem to attribute agency and intention more readily to the causes of negative outcomes than to the causes of positive outcomes (Hamlin & Baron, 2014; Morewedge 2009). These findings together with our data suggest that there are separate cognitive processes at work for the attribution of intention when the presence of an agent is uncertain and for the attribution of intention when the presence of an agent is certain. In the former case (as the work of Hamlin and Morewedge suggests), the presence of a negative outcome is a cue that an agent is present and that the negative outcome was brought about intentionally. In the latter case (as the work reported here suggests), when the presence of an agent has been established and a negative outcome is observed, the Good Intention Prior is applied and the negative outcome is not seen as intentional (given that there is a good effect that can plausibly be the agent’s goal instead).
Second, on the face of things, the side-effect effect seems to be at odds with a good intention prior. The side-effect effect is the phenomenon whereby disavowed negative side-effects are seen as more intentional than their positive counterparts (e.g. Knobe, 2003; Leslie, Knobe & Cohen, 2006). Our findings seem to fit the reverse pattern, that positive effects are seen as more intentional. There are two key differences between our phenomenon and the side-effect effect that eliminates the seeming contradiction. First, in side-effect effect style cases, the protagonist in the story disavows one of the effects, that is, he explicitly declares that he does not care about it. While it might seem that this declaration is just a way of indicating that an effect is a side-effect (counterfactually irrelevant to the agent’s action plan), subsequent studies have convincingly shown that such a statement indicates a certain kind of added intention in the case of the negative effect and not the positive effect (Uttich & Lombrizo, 2010; Guglielmo & Malle, 2010; Nanay, 2010; Sripada, 2009). Our finding suggests that effects are seen as intended when no other intention information is present; side-effect effect style cases apparently provide information concerning the protagonist’s attitude towards the effects.

The second difference between the side-effect effect phenomenon and ours is that the side-effect effect seems to be most robust when subjects are asked if the protagonist acted “intentionally”; the effect is much weaker when subjects are asked if it was the agent’s “intention” to bring about the the good and bad effects (Knobe, 2004). As discussed in the introduction, our phenomenon concerns goal-attribution (acting with an intention) whereas the side-effect phenomenon seems to be about acting intentionally (cf. Premack, 1990; Searle, 1983).
A new set of questions about the Good Intention Prior arises from our results. For example, how much (and what quality) of countervailing information is required to override the prior? If countervailing intention information is provided in the stimulus for one of two effects of an agent’s action, how do subjects represent the agent’s intention towards the other effect? Does the information provided by the prior remain in place for that other effect? Does the disruption of one part of the prior mean that the prior is entirely silenced?

Finally, is this prior specific to the moral domain, or does it span other normative or evaluative domains (such as aesthetics, epistemology, economics, and so forth)? The side-effect effect phenomenon does not seem to be restricted to the moral domain, but instead spans other evaluative domains (Knobe & Mendlow, 2004; Machery, 2008; Uttich & Lombrozo, 2010; Rakoczy et al., 2015). The side-effect effect is also a phenomenon concerning intending according to norms, so this potentially suggests that the good intention prior effect might also span other evaluative domains.

5. Conclusion

Our data suggests that a good intention prior develops by 3-years of age at the latest and persists in adult cognition. This prior has likewise been canonized as a “presumption of innocence” in international human rights law (General assembly resolution 217(III)). While it seems like this prior could potentially be learned in early development (as we have suggested above), it also seems like a good candidate to be built into core cognition, a kind of cognitive constraint that allows morally relevant intention structures to be parsed early on (cf. Spelke, Bernier & Skerry, 2013). Indeed, Mikhail (2012) points out a long philosophical tradition describing the innate instinct to assume that those around us
“pursue good and avoid evil” (e.g. Hume 1978/1740: 438; Aquinas 1988/1274: 49). After all, without a presumption of innocence, communication would break down (under constant suspicion of deception; Grice, 1989), economies couldn’t function (under constant suspicion of fraud; Henrich et al., 2007), and most other basic social interactions would be rendered futile. That we think others are up to good ends binds our social world together.
III. Preschoolers’ Use of Moral Rules and Representations

Taking a grammar to be a system of rules that provides representations of sound and meaning (among others), their specific character to be determined as research progresses, our task is to discover the representations that appear and the rules operating on them and relating them; and more important, to discover the system of universal grammar that provides the basis on which they develop.

-Noam Chomsky (1980, pg 8)

Recent research in moral psychology has suggested that adults use rules that are not consciously accessible to make moral judgments (Mikhail, 2011; Greene, 2013; Mallon & Nichols, 2010; Cushman, Young & Hauser, 2006; Bonnefon, Shariff & Rahwan, 2016). These rules seem to drive strong intuitive judgments that, upon reflection, we may find hard to explain. Yet even when the inexplicability of our judgments are drawn to our attention, we are sometimes unwilling to let them go (Mikhail, 2007; cf. Haidt, 2001). Is it possible that the unconscious moral rules driving our judgments are learned – either didactically or implicitly through years of socialization? Or are some of these unconscious moral rules more core, shaping how we process the moral world from a young age (Spelke, Bernier & Skerry, 2013)? We argue that already by preschool, children seem to make moral judgments on complex cases of welfare tradeoffs by using abstract rules that go well beyond the sorts of rules they are taught by parents and teachers. If these rules are learned implicitly, they require powerful learning mechanisms that operate early on in life.

The terms of the moral rules that preschoolers use can hint at the terms in which the moral world is represented (Mikhail, 2011; cf. Chomsky, 1980). If a rule concerns, for instance, the permissibility of pushing, then in order to apply the rule, a cognizer must be able to represent whether a push has happened or not. We argue that the rules used by preschool-aged children are not written exclusively in terms of readily observable
features of the moral world (eg, hitting and taking), but that involve more abstract terms (eg, utility structures and intentional action plans).

In Experiment 1 we present evidence that suggests that preschoolers use the “means principle” in their moral evaluations, treating harm caused as a means as worse than harm caused as a foreseen side-effect. This suggests that our subjects are representing the intentional action plans of moral agents with a fine enough granularity to make the means/side-effect distinction. Experiment 2 suggests that evaluations on the grounds of the means/side-effect distinction may be restricted to the moral domain rather than being a domain-general rule.

In Experiment 3 we present evidence that suggests that preschoolers use the implied consent principle, permitting an otherwise morally problematic action if the victim would consent to the violation. This suggests that our subjects are representing the utility structures of moral patients, weighing the gravity of specific harms and using that assessment to judge moral actions. Experiment 4 suggests that preschoolers do this even when the two competing harms come from disparate domains.

**Experiment 1**

In Experiment 1, we looked at whether preschoolers treat harm caused as a means as morally worse than harm caused as a foreseen side-effect.

**Methods**

**Pretesting.** Subjects in all experiments reported here were tested individually in quiet locations in their preschools or in the lab. Subjects in all experiments were given the
following training and prescreen. Subjects were trained on a Likert scale with X’s on one end and stars on the other, with points corresponding to “really bad”, “a little bad”, “just OK”, “a little good”, “really good”. Subjects were told stories in which a simple morally good or bad action took place. Children were asked to issue a moral judgment of the action (“Should he/she have done that?”) and were asked to rate the action on the Likert Scale. Only children who expressed competence making simple moral assessments and using the scale to describe moral behavior were tested further (see Appendix A).

Procedure. Subjects were told a story accompanied by an animation about a mean squirrel in a park that was about to eat five children’s cookies. Another girl who does not have a cookie, the protagonist, sees what is about to happen and chooses to prevent the squirrel from eating the five children’s cookies by causing the squirrel to eat another child’s cookie (the bystander), who was previously unthreatened.

Subjects were randomly assigned to one of two conditions: Means or Side-effect. In the Means Condition, the protagonist waves to the bystander and calls for her to come over. When she arrives, the protagonist takes her cookie and feeds it to the squirrel, giving the other five children time to eat their cookies. This condition involves harm as a means because the harm caused (the bystander’s cookie being eaten by the squirrel) is a necessary part of the action plan of the agent; if that effect had not occurred, the goal (saving the five kids’ cookies) would not have been brought to fruition. In the Side-Effect Condition, the bystander is carrying around a long gate with her. The protagonist waves to the bystander and calls for her to come over. The bystander approaches and her path intersects the path of the squirrel; the gate she is carrying blocks the squirrel, preventing it from eating the cookies of the five children. Meanwhile, the squirrel eats
the bystander’s cookie. This condition involves harm as a side-effect because the harm caused is not a necessary part of the action plan of the agent; the squirrel’s eating of the bystander’s cookie is incidental of the plan of having the bystander’s gate block the squirrel. If the bystander’s cookie had not been eaten at all, the goal would still have been achieved. In both conditions, emphasis was placed on the fact that the protagonist knows what is going to happen to the bystander when she calls her over. (See Appendix C for exact script of each story for this and subsequent experiments.)

In this and all subsequent studies, subjects were told the story with the accompanying animation twice. On the second telling, subjects were asked a series of questions about the events of the story. If subjects failed to remember critical facts about the story, they were excluded from analysis (see Appendix C for list of control questions and exclusion criteria for each Experiment).

In this study, subjects were asked two test questions, the first concerning the permissibility of the protagonist’s action (“Should she have done that?”), the second asking for a Likert scale rating of the agent’s action (“Was what Sally did good, bad, or just OK?”).

Figure 3.1: Selected scenes from Side-Effect and Means Conditions (Experiment 1) and Means without Stealing (Experiment 1b).
Subjects. Eighty four children between the ages of 3.9 and 6.0 years participated in the study (M=4.82 years; SD=.51), 40 of which were girls. Forty-two children received the Means Condition (M = 4.79 years; SD = .58), 21 of which were girls. Forty-two children received the Side-Effect Condition (M = 4.85 years; SD = .4), 19 of which were girls. Twenty additional children began the study but were excluded, 13 for failing scale training, 6 for failing control questions, and 1 for not cooperating. We stopped data collection after getting data from 84 children in order to reach sample sizes that have been previous used for studies with a similar methodology (cf. Saunders, 2014).

Results

In the Means Condition, 11.9% of subjects (5 out of 42) responded “yes” to the question “Should she have done that?” (binomial test, N=42, x=5, p<.0001, two-tailed). In the Side-Effect Condition, 66.7% of subjects (28 out of 42) said yes to the “should” question (binomial test, N=42, x=14, p=.02, two-tailed).

Subjects were significantly more likely to answer yes to the “should” question in Side-Effect Condition than the Means Condition, Upton’s $\chi^2(1, n = 84) = 26.1, p <.0001$, two-tailed.
Figure 3.2: Subjects’ responses to the question, “Should she have done that?”

Likert scale ratings (here, and in all subsequent experiments) were scored as follows: really bad = -2; a little bad = -1; just OK = 0; a little good = 1; really good = 2. Subjects in the Means Condition rated the action as bad (M= -.95; SE=.19), while subjects in the Side-Effect condition rated the action as Just-OK (M=.12; SE=.22). Subjects in the Means Condition rated the agent’s action as significantly worse than did subjects in the Side-Effect Condition (Means: M = -.95, SE = .19; Side-Effect: M = .12, SE = .22; independent-sample t-test, t(82)=3.68, p=.0004, two-tailed).

Figure 3.3: Subjects’ ratings of the action of the protagonist. Error bars shows standard error of the mean.
Experiment 1b

The results of Experiment 1 suggest that preschoolers treat harm caused as a means as morally worse than harm caused as a foreseen side-effect. However, an alternate explanation of the findings in Experiment 1 is possible. In the Means Condition, in addition to causing harm as a means, the protagonist steals and destroys the bystander’s property. In the Side-Effect Condition, the protagonist simply waves at the bystander. Experiment 1b was designed to test whether this difference generated the effect observed in Experiment 1. Experiment 1b involves a protagonist causing harm as a means, but without the additional morally problematic act of stealing.

Methods

Methods were the same as for Experiment 1 with the following modifications. As the squirrel approaches the five kids’ cookies, the protagonist calls to the bystander to come over, knowing that the bystander will intersect the path of the squirrel. The squirrel eats the bystander’s cookie, allowing the five kids to eat their cookies in the meantime.

Subjects. Forty-two children between the ages of 4.1 and 6.2 years participated in the study (M=5.07 years; SD=.47), 20 of which were girls. We stopped data collection after getting data from the same number of participants as were in each condition of Experiment 1.

Results
26.2% of subjects (11 out of 42) responded “yes” to the question “Should she have done that?” (binomial test, N=42, x=11, p=.001, two-tailed).

There was no significant difference between subjects’ responses to the Means No Stealing case from this experiment and the Means Stealing case from Experiment 1, Upton’s $\chi^2(1, n = 84) = 2.7, p = .1$, two-tailed, though there is an effect trending in the direction of the Means Stealing case being judged less permissible. There was a significant difference between the Side-Effect case in Experiment 1 and Means No Stealing, Upton’s $\chi^2 (1, n = 84) = 13.7, p = .0002$, two-tailed.

![Graph showing responses to the question, “Should she have done that?” from Study 1 and Study 1b.]

Figure 3.4: Subjects’ responses to the question, “Should she have done that?” from Study 1 (first two bars) and Study 1b (last bar).

Subjects in Experiment 1b rated the agent’s action as bad ($M = -.55; SE = .18$). When all ratings are considered from Experiment 1 and Experiment 1b, analysis of variance revealed a significant difference between the conditions, $F(2, 123) = 7.53$, $\eta^2=.11, p=.0008$. Planned pairwise comparisons revealed that there was no significant difference between Means No Stealing and Means Stealing (independent-sample t-test, two-tailed, $t(82) = 1.57, p=.12$). There was a significant difference between the Means
No Stealing and Side-Effect (independent-sample t-test, $t(82) = 3.66$, $p=.0004$, two-tailed).

![Graph showing ratings of action of the protagonist in Experiment 1 and 1b.](image)

Figure 3.5: Subjects’ ratings of the action of the protagonist Experiment 1 (first two bars) and Experiment 1b (last bar). Error bars shows standard error of the mean.

**Discussion**

Results from Experiment 1 and 1b suggest that preschoolers treat harm caused as a means as morally worse than harm caused as a foreseen side-effect and that this effect is not contingent on the presence of another morally problematic action, such as stealing.

**Experiment 2**

In Experiment 2, we tested whether preschoolers would use the means/side-effect distinction in a non-moral scenario.

**Methods**

Subjects were told a story accompanied by an animation about a rock rolling down a hill that is about to squish five apples. A protagonist intervenes in a manner that destroys
a lone apple that was not originally in the path of the rock, while preventing the five apples from being destroyed.

Subjects were randomly assigned to one of two conditions: Means or Side-effect. In the Means Condition, the protagonist throws the lone apple at the rock, causing the rock to change course and avoiding squishing the five apples. This condition involves squishing the apple as a means; the collision of the apple with the rock is a necessary part of the action plan of the agent and if that effect had not occurred, the goal (preventing the five apples from being squished) would not have come about. In the Side-Effect Condition, the protagonist gets a ramp and places it in front of the rolling rock. The rock rolls down the ramp, avoiding the five apples and instead squishing the lone apple. This condition involves squishing the apple as a side-effect because the collision of the rock and the apple is not a necessary part of the action plan of the agent. If the lone apple had not been there at all, the goal would still have been achieved. In both conditions, the protagonist leaves the scene to go home after intervening without the apples.

Subjects were asked three test questions, the first concerning the permissibility of the protagonist’s action (“Should he have done that?”), the second asking for a Likert scale rating of the agent’s action (“Was what Billy did good, bad, or just OK?”), and the third concerning the intention of the agent (“Did Billy squish the apple on purpose?”).
Figure 3.6: Selected scenes from Side-effect and Means condition of Experiment 2.

(Arrows added for emphasis.)

Subjects. Ninety-seven children between the ages of 3.88 and 5.67 years participated in the study (M=4.82 years; SD=.41), 46 of which were girls. Forty-nine children received the Means Condition (M = 4.84 years; SD = .43), 17 of which were girls. Forty-eight children received the Side-Effect Condition (M = 4.80 years; SD = .39), 29 of which were girls. Thirty-one additional children began the study but were excluded, 24 for failing scale training, 3 for failing control questions, and 4 for refusing to answer all test questions. We planned to stop data collection when data was collected for 42 subjects in each condition to match sample sizes to Experiment 1, however data was lost and then recovered, so sample sizes were larger than expected.

Results

In the Means Condition, 69.4% of subjects (34 out of 49) responded “yes” to the “should” question (binomial test, N=49, x=15, p=.0047, two-tailed). In the Side-Effect
Condition, 83.3% of subjects (40 out of 48) responded “yes” (binomial test, N=48, x=8, p<.0001, two-tailed).

There was no significant difference between subjects answers to the “should” question in Side-Effect Condition and the Means Condition, Upton’s χ²(1, n = 97)=2.58, p=.11, two-tailed.

![Figure 3.7: Subjects’ responses to the question, “Should he have done that?”](image)

Subjects in both conditions rated the action as somewhat good (Means: M= .35; SE=.18; Side-Effect: M=.73 ; SE=.17). There was no significant difference between ratings in the Means and Side-Effect Conditions (independent samples t-test, t(95)=1.53, p=.13, two-tailed).
Preschoolers do not seem to draw a distinction between the Means and Side-Effect conditions on the evaluative questions. However, they do acknowledge that there is a difference in the intention of the agent. In the Means Condition, 49.0% of subjects (24 out of 49) judged the squishing of the one apple intended, responding “yes” to the “on purpose” question (binomial test, N=49, x=22, p=.5, two-tailed). In the Side-Effect Condition, 18.6% of subjects (9 out of 48) responded “yes” (binomial test, N=48, x=9, p<.0001, two-tailed). Subjects were significantly more likely to answer “yes” to the “on purpose” question in the Means Condition than in the Side-Effect Condition, Upton’s \( \chi^2(1, n = 97) = 9.77, p = .0018, \) two-tailed.
Figure 3.9: Subjects’ responses to the question, “Did Billy squish the apple on purpose?”

**Discussion**

In Experiment 2, subjects seemed to be representing action plans in the way they did in Experiment 1 (with enough subtly to recognize that one of the cases involved squishing the apple as a means and one involved doing so as a side-effect), though that fact did not influence their evaluations of the protagonist’s action. This suggests that the means/side-effect distinction may be constrained to evaluations of moral actions – or at least does not operate in an entirely domain-general way. It is notable that even in the Means Condition, only about half of subjects judged that the apple was squished “on purpose.” This may be the case because the squishing of the apple is not strictly necessary for the protagonist’s goal to be completed, only (the closely related event of) the collision of the apple with the rock such that the rock veers off course. Future work should investigate under what circumstances intention ascription is impacted by the perceived granularity of event segmentation.
Experiment 3

In this experiment, we looked at whether an abstract rule that preschoolers have not been explicitly taught can override a rule they typically learn from parents and teachers.

Methods

Subjects were told a story accompanied by an animation about a girl (the bystander) playing with five beach ball. Subjects were randomly assigned to one of two conditions: Implied Consent or Baseline. In the Implied Consent condition, a mean crab enters the scene and is about to pop all the beach balls. The protagonist takes one ball and throws it, causing the crab to chase the ball and leave. In the Baseline Condition, the protagonist simply takes one ball away from the bystander and throws it.

Subjects were asked two test questions, the first concerning the permissibility of the protagonist’s action (“Should she have done that?”) and the second asking for a Likert scale rating of the agent’s action (“Was what Jane did good, bad, or just OK?”). In the Implied Consent Condition, another question concerning the relative value of the impending harms was asked after the other two test questions (“What would be better for this girl? If the crabs broke her balls, or if Jane took her ball?”). The order of answer choices for the latter question were counterbalanced.
Implied Consent Condition (Experiment 3)  Implied Consent Condition (Experiment 4)

Figure 3.10: Scenes from the Implied Consent Conditions of Experiment 3 and 4.
(Arrows added for emphasis.)

**Subjects.** Forty-six children between the ages of 3.45 and 5.68 years participated in the study (M=4.79 years; SD=.53), 26 of which were girls. Thirty-two children received the Implied Consent Condition (M = 4.81 years; SD = .44), 19 of which were girls. Fourteen children received the Baseline Condition (M = 4.76 years; SD = .7), 7 of which were girls. Twelve additional children began the study but were excluded, 10 for failing scale training, 2 for failing to refusing to answer all the questions.

**Results**

In the Baseline, 28.6% of subjects (4 out of 14) responded “yes” to the “should” question (binomial test, N=14, x=4, p=.09, two-tailed). In the Implied Consent Condition, 62.5% of subjects (20 out of 32) responded “yes” (binomial test, N=32, x=12, p=.11, two-tailed).
There was a significant difference between subjects’ answers to the “should” question in the Baseline and the Implied Consent Conditions, Fischer’s Exact Test, $p=.054$, two-tailed.

Figure 3.11: Subjects’ responses to the question, “Should she have done that?”

Subjects in the Baseline Condition rated the action as bad ($M=-1.21; SE=.30$), while subjects in the Implied Consent condition rated the action as somewhat positive ($M=.59; SE=.25$). Ratings in the Baseline Condition were significantly worse than ratings in the Implied Consent Condition (independent samples t-test, $t(44)=4.18$, $p=.00014$, two-tailed).
Each subject in the Implied Consent Condition was asked to make a relative value assessment about the harms facing the victim. Of the 32 subjects, 26 responded that it would be better for the girl to lose her ball than for all the balls to be popped by the crab; the remaining 6 subjects made the opposite assessment (presumably judging that to be intentionally harmed by a peer is worse than to be the victim of a “natural disaster”; cf. philosophical proponents of relational equalitarianism in opposition to luck egalitarianism, Anderson, 1999). Analysis of the “should” question and action ratings was conducted based on subjects’ value assessments. Of the subjects who said that losing one ball would be better, 73.1% of subjects (19 out of 26) responded “yes” to the “should” question (binomial test, N=26, x=7, p=.01, two-tailed). Of the subjects that said that being bitten would be better, 16.7% of subjects (1 out of 6) judged the case permissible (binomial test, N=6, x=1, p=.1, two-tailed).
There was a significant difference between subjects answers to the “should” question when subjects were divided based on their answers to the “Which would be better” question, Fischer’s Exact Test, \( p = .019 \), two-tailed.

![Bar chart showing responses to question]

Figure 3.13: Subjects’ responses to the question, “Should she have done that?” The left bar represents the responses of the subjects that said that losing one ball would be better than losing all the balls. The right bar represents the responses of subjects that said losing all the balls would be better than losing one ball.

Subjects who judged that losing one ball would be better, rated the action of the agent as somewhat good (\( M = .73; SE = .27 \)), while subjects who judged that being bitten would be better rated the action of agent as just OK (\( M = 0.0; SE = .68 \)). There was no significant difference between ratings in the two groups (independent samples t-test, \( t(30) = 1.13, p = .26 \), two-tailed).
Discussion

The results of Experiment 3 suggest that children are not rotely applying rules they learn from caregivers, but instead apply more abstract versions of those rules that take into account the utility structures of moral patients. Subjects assess that taking something from another child is permissible when that child would be better off due to the stealing (and not permissible when the child would not be better off) – though this iteration of the “don’t take things that don’t belong to you” rule was perhaps never uttered by these subjects’ caregivers. Future work should investigate whether it is the subjects’ own view of what counts as being “better off”, her judgment of the view of the moral patient, or her judgment of the protagonist’s judgment of the patient’s view that is driving her assessment of the protagonist’s action.
Experiment 4

The goal of Experiment 4 was to replicate the findings of Experiment 3 in a context that involved comparing harms across different domains (property violation and battery).

Methods

Subjects were told a story accompanied by an animation about a girl (the bystander) playing with a ball. Subjects were randomly assigned to one of two conditions: Implied Consent or Baseline. In the Implied Consent condition, a mean dog enters the scene and is about to bite the bystander. The protagonist takes the ball away from the bystander and throws it away, causing the dog to chase the ball and leave. In the Baseline Condition, the protagonist simply takes the ball away from the bystander and throws it.

Subjects were asked two test questions, the first concerning the permissibility of the protagonist’s action (“Should he have done that?”) and the second asking for a Likert scale rating of the agent’s action (“Was what Jane did good, bad, or just OK?”). In the Implied Consent Condition, another question concerning the relative value of the impending harms was asked before the other two test questions (“What would be better for this girl? If the dog bit her, or if Jane took her ball?”). The order of answer choices for the latter question were counterbalanced. This test question was asked before the other two questions to control for the possibility that answering the questions concerning permissibility and action evaluation first (as in Experiment 3) encouraged them to give certain answer to the value question.
Subjects. Thirty-six children between the ages of 3.47 and 5.61 years participated in the study (M=4.50 years; SD=.52), 23 of which were girls. Twenty-four children received the Implied Consent Condition (M = 4.47 years; SD = .45), 14 of which were girls. Twelve children received the Baseline Condition (M = 4.57 years; SD = .65), 9 of which were girls. Three additional children began the study but were excluded for failing to answer all the test questions.

Results

In the Baseline, 8.3% of subjects (1 out of 12) responded “yes” to the “should” question (binomial test, N=12, x=1, p=.003, two-tailed). In the Implied Consent Condition, 66.7% of subjects (16 out of 24) responded “yes” (binomial test, N=24, x=16, p=.08, two-tailed).

There was a significant difference between subjects’ answers to the “should” question in the Baseline and the Implied Consent Conditions, Fischer’s Exact Test, p=.0013, two-tailed.

Figure 3.15: Subjects’ responses to the question, “Should she have done that?”
Subjects in the Baseline Condition rated the action as bad (M=-1.33; SE=.19), while subjects in the Implied Consent condition rated the action as just OK (M=.29; SE=.31). Ratings in the Baseline Condition were significantly worse than ratings in the Implied Consent Condition (independent samples t-test, t(34)=3.53, p=.0012, two-tailed).

Figure 3.16: Subjects’ ratings of the action of the protagonist. Error bars show standard error of the mean.

Of the 24 subjects, 17 responded that it would be better for the girl to lose her ball than to be bitten; the remaining 7 subjects made the opposite assessment. Analysis of the “should” question and action ratings was conducted based on subjects’ value assessments. Of the subjects who said that losing the ball would be better, 88.2% of subjects (15 out of 17) judged the case responded “yes” to the “should” question (binomial test, N=17, x=2, p=.001, two-tailed). Of the subjects that said that being bitten
would be better, 14.3% of subjects (1 out of 7) judged the case permissible (binomial test, N=7, x=1, p=.06, two-tailed).

There was a significant difference between subjects answers to the “should” question when subjects were divided based on their answers to the “Which would be better” question, Fischer’s Exact Test, p=.0013, two-tailed.

![Figure 3.17: Subjects’ responses to the question, “Should she have done that?”](image)

Figure 3.17: Subjects’ responses to the question, “Should she have done that?” The left bar represents the responses of the subjects that said that losing the ball would be better than being bitten by the dog. The right bar represents the responses of subjects that said being bitten would be better than losing the ball.

Subjects who judged that losing the ball would be better, rated the action of the agent as somewhat good (M=.47; SE=.35), while subjects who judged that being bitten would be better rated the action of agent as just OK (M=-.14; SE=.63). There was no significant difference between ratings of the two groups (independent samples t-test, t(22)=.90, p=.38, two-tailed).
Figure 3.18: Subjects’ ratings of the action of the protagonist. Error bars show standard error of the mean. The left bar represents the responses of the subjects that said that losing the ball would be better than being bitten by the dog. The right bar represents the responses of subjects that said being bitten would be better than losing the ball.

**Discussion**

Experiment 4 replicates the main finding of Experiment 3: subjects assess that taking something from another child is permissible when that child would be better off due to the stealing (and not permissible when the child would not be better off). Preschoolers make this assessment comparing harm in the domain of property to harm in the domain of battery.

**General Discussion**

Our results suggest that by preschool age, children are using abstract rules to make moral judgments that operate over factors that are not directly observable.

Our results concerning the means principle are consistent with previous work highlighting the importance of mental state inference in socio-moral judgment in infants and preschoolers. Recent work suggests that by the first year of life, infants evaluate
agents based on their goals to help and hinder others (Hamlin, Wynn & Bloom, 2007; Hamlin, Wynn, Bloom, Mahajan, 2011) which cannot be explained by infants’ sensitivity to behavioral cues (Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013). By preschool age, children can use their representation of others’ goals to make moral judgments (Cushman, Sheketoff, Wharton, & Carey, 2013; Baird & Astington, 2004; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011) using simple moral rules (eg: intending to bring about a bad goal is not permissible).

Our work extends these studies in two ways. First, previous studies have pitted agents with good goals against agents with bad goals. In our studies, agents in the Side-Effect and Means Conditions have the same goal (to save five children); the difference in the conditions lies in the agents’ step-wise action plans. This suggests that children are representing more fine-grained mental state information than simply an agent’s goal state. Second, our studies imply that children are not limited to moral rules that operate over agents’ goals, but that they use a more subtle moral rule (the means principle) that operates over action plans. Moreover, this rule seems to swamp any effect of a rule that operates over concrete properties of the moral world (eg: an observed act of stealing).

Our work concerning the implied consent principle extends previous work suggesting that young children can use mitigating factors to ameliorate the harshness of their judgments of prima facie wrongs. Work on this topic has mostly looked at children’s judgments of prososial lie-telling to spare someone’s feelings (eg, saying they liked a gift they were disappointed by; Bussey, 1999; Walper and Valtin, 1992; Xu, Bao, Fu, Talwar, & Lee, 2010), finding that young children rate prosocial lying as better than self-serving lying. These results are suggestive, though they could be explained by the
fact that children gradually learn an exception to the prohibition against lie-telling exists in cases where the socio-conventional norms of politeness are at stake. Further work in this vein has suggested that children judge a protagonist’s harming of a would-be transgressor as better than similar harm done for selfish purposes (Rule, Nesdale, & McAra, 2974; Darley, Klosson & Zanna, 1978; Jambon & Smetana, 2013). However, in those studies it is possible children permit the harm using a rule that “the transgressor sacrifices rights he or she would otherwise have by the intentional breach of another's rights” (Greenawalt, 1983, pg. 454) rather than weighing the harms at stake.

Our work reveals a more fundamental ability to represent and compare a common harm-based moral currency across moral violations from different domains. Moreover, our work suggests that preschoolers use a rule that permits a moral wrong against a victim if that victim emerges from the scenario better off. Note that this seems to be a kind of utilitarianism that is applied when harms and benefits are directed towards one individual only; Experiment 1 rules out that preschoolers are simply doing utilitarian calculus in all morally charged scenarios.

**Conclusion**

Together, these studies suggest the use of unconscious abstract rules by preschool aged-children, which operate over representations of intentional action plans and utility structures. The intuitive moral judgments given by adults may have a very early basis, indeed.
IV. The Role of Choice in Moral Judgment

Introduction

How do we decide if an action is morally permissible? Leading theories of cognitive moral psychology typically focus on the roles of intention ascription and outcome assessment and try to tease out how these two cognitive processes interact on judgments of moral wrongness and blame (e.g., Cushman, Sheketoff, Wharton & Carey 2013; Greene 2013; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011).

While outcomes and intentions clearly play a central role in moral judgment, we suggest that a picture of moral cognition that considers only outcomes and intentions lacks a key feature: the consideration of morally preferable alternative actions (Mikhail, 2011). If permissibility judgments are based on calculations of net consequences coupled with consideration of the agent’s intention, then as long as the agent intends the good effects and the outcome is net positive, the action should be permissible.

However, a moment’s reflection suggests that this cannot be the full account. Recent attention to excessive use of force by police officers bring this concern into the spotlight. Even when an officer incapacitates a dangerous criminal, preventing potentially significant harm to innocent bystanders, public outrage can result when the officer is seen as having used excessive force against the person being apprehended, that is, foregoing a morally better alternative that is pareto optimal – one that would have resulted in the same good effects with fewer bad ones. Indeed, the Supreme Court of the United States has ruled that excessive police force infringes on Fourth Amendment Rights (Harmon, 2008) and victims of excessive police force can bring legal suits against the officers that have apprehended them (Scott v. Harris, 2006).
As this case-study suggests, in addition to considering morally relevant outcomes and intentions, we suggest that we routinely consider whether morally better alternatives were available when making moral judgments.

Alternately, it is possible that considering the choice set available to a moral agent plays only an *indirect* role in moral judgment. It is possible that assessment of the choices available to an agent impacts our assessment of the agent’s intention, which then impacts our moral judgment. For example, in cases of police brutality, it is possible that the excessive use of force indicates that the officer *intended to harm* the victim in addition to protect citizens. If this is the case, then our moral outrage at these cases could be explained by the fact that choice informs us about intention (which impacts moral judgment) and not that choice has an independent effect on moral judgment.

We present two studies which suggest that preschoolers and adults judge that some otherwise permissible actions are forbidden when morally preferable alternatives are available. Our results in both studies also suggest that the role of choice cannot be reduced to providing information about an agent’s intention.

**Experiment 1**

In Experiment 1, we looked at whether preschoolers take into account the available alternatives when making a moral judgment.

**Methods**

**Pretesting.** Subjects in all experiments reported here were tested individually in quiet locations in their preschools or in the lab. Subjects were trained on a Likert scale with X’s on one end and stars on the other, with points corresponding to “really bad”, “a little
bad”, “just OK”, “a little good”, “really good”. Subjects were told stories in which a simple morally good or bad action took place. Children were asked to issue a moral judgment of the action (“Should he/she have done that?”) and were asked to rate the action on the Likert Scale. Only children who expressed competence making simple moral assessments and using the scale to describe moral behavior were tested further (see Appendix A).

Procedure. Subjects were told a story accompanied by an animation about a mean squirrel in a park that was about to eat three children’s cookies. Another girl who does not have a cookie, the protagonist, sees what is about to happen and thinks about what to do. She then chooses to prevent the squirrel from eating the three children’s cookies by putting up a gate.

Subjects were randomly assigned to one of three conditions: Save Some, Better Alternative, and Save All. In the Save Some Condition, the protagonist puts up her gate in such a way that the three children are blocked from the squirrel, but the squirrel is redirected to eat the cookie of a fourth child (the bystander) who was heretofore unthreatened. In the Better Alternative Condition, the bystander considers putting up her gate in such a way that all four children in the scene will be protected and then considers putting it in the “Save Some position”; she ultimately puts it in the “Save Some” position. In the Save All Condition, the protagonist considers both gate positions, and puts up her gate in the orientation that saves all four children. In the latter two conditions, the order of gate presentation was counterbalanced (Figure 1; see Appendix D for exact script of each story.)
Figure 4.1: Selected scenes from each condition. In the Save Some Condition (left), the protagonist chooses to put up her gate such that some of the children’s cookies are saved. No other options are mentioned. In the Better Alternative Available Condition (middle), the protagonist has the option of putting her gate in the “Save All” position (light red bar), but chooses to put her gate in the “Save Some” position (dark red bar). In the Save All Condition (right), the protagonist chooses to put up her gate such that all of the children’s cookies are saved. No other options are mentioned.
Subjects were told the story with the accompanying animation twice. On the second telling, subjects were asked a series of questions about the events of the story. If subjects failed to remember critical facts about the story, they were excluded from analysis (see Appendix D for list of control questions and exclusion criteria).

Subjects were asked three test questions, the first concerning the permissibility of the protagonist’s action (“Should she have done that?”), the second asking for a Likert scale rating of the agent’s action (“Was what Sally did good, bad, or just OK?”), and the third concerning Sally’s intention towards the bystander (“Did Sally make this girl sad/happy on purpose?”). Some subjects were asked a fourth question about the protagonist’s intention (“Did Sally make this group of kids happy on purpose?”) but few children attended to the stimuli long enough to answer that question, so it was dropped from analysis.

Some subjects spontaneously offered suggestions of an action the protagonist could have taken that would have led to a morally better outcome than the one that they saw (eg, “She should have chased the squirrel away.”). A spontaneous utterance was coded as a better alternative if the subject 1) refers to something other than what actually happened and 2) describes an action (e.g., “put up the gate here”) or an outcome (e.g. “make everyone happy”) that would lead to a better outcome than the one that actually takes place.

Subjects. One hundred nineteen children between the ages of 3.7 and 6.2 years participated in the study (M=4.9 years; SD=.47), 60 of which were girls. Forty-seven children received the Save Some Condition (M = 5.0 years; SD = .42), 28 of which were
girls. Forty children received the Better Alternative Available Condition (M = 4.9 years; SD = .54), 22 of which were girls. Thirty-two children received the Save All Condition (M = 4.8 years; SD = .46), 10 of which were girls. Fifty additional children began the study but were excluded, 17 for failing scale training, 29 for failing control questions, and 4 for not answering all test questions.

**Results**

In the Save Some Condition, 76.6% of subjects (36 out of 47) responded yes to the question “Should she have done that?” (binomial test, N=47, x=11, p=.0002, two-tailed). In the Better Alternative Available Condition, 27.5% of subjects (11 out of 40) said yes to the “should” question (binomial test, N=40, x=11, p=.003, two-tailed). In the Save All Condition, 93.8% of subjects (30 out of 32) said yes to the “should” question (binomial test, N=32, x=2, p<.0001, two-tailed).

Subjects were significantly more likely to answer yes to the “should” question in the Save Some Condition than the Better Alternative Condition (Upton’s χ²(1, n = 87) = 20.7, p <.0001, two-tailed) and in the Save All Condition than the Better Alternative Condition (Upton’s χ²(1, n = 72) = 31.4, p <.0001, two-tailed).
Figure 4.2: Subjects’ responses to the question, “Should she have done that?”

Likert scale ratings were scored as follows: Really Bad = -2; A Little Bad = -1; Just OK = 0; A Little Good = 1; Really Good = 2. Subjects in the Save Some and Better Alternative Conditions rated the action as Just OK (Save Some: M= .21; SE=.20; Better Alternative: M = -.40; SE = .20). Subjects in the Save All condition rated the action as Good (M=1.25; SE=.19).

Analysis of variance revealed a significant difference between the conditions, \( F(2, 119) = 15.6, p<.0001 \). Ratings in the Better Alternative Condition were significantly worse than ratings in the Save Some Condition (independent-sample t-test, \( t(85)=2.16, p=.03, \text{two-tailed} \)) and significantly worse than ratings in the Save All Condition (independent-sample t-test, \( t(70)=5.95, p<.0001, \text{two-tailed} \)).
Figure 4.3: Subjects’ ratings of the action of the protagonist. Error bars show standard error of the mean.

In the Save Some Condition, 46.8% of subjects (22 out of 47) responded yes to the question “Did she make that one happy/kid sad on purpose?” (binomial test, N=47, x=22, p=.39, two-tailed). In the Better Alternative Available Condition, 55.0% of subjects (22 out of 40) said yes to the “on purpose” question (binomial test, N=40, x=18, p=.32, two-tailed). In the Save All Condition, 59.4% of subjects (19 out of 32) said yes to the question “Did she make that one kid happy on purpose?” (binomial test, N=32, x=13, p=.19, two-tailed).

There were no significant differences in subjects’ answers to the “on purpose” question in the Save Some Condition compared to the Better Alternative Condition, Upton’s $\chi^2(1, n = 87) = .57, p = .45$, two-tailed. The Bayes Factor provides moderate support for the null hypothesis (BF = 2.9). There were also no significant differences in subjects’ answers to the “on purpose” question in the Save All Condition compared to
the Better Alternative Condition, Upton’s $\chi^2(1, n = 72)= .14, p = .71$, two-tailed. The Bayes Factor provides moderate support for the null hypothesis (BF = 3.3).

![Graph showing responses to the question, “Did she make the one kid happy/sad on purpose?”](image)

Figure 4.4: Subjects’ responses to the question, “Did she make the one kid happy/sad on purpose?”

While there is no difference in intention judgments across conditions, is there an effect of intention judgments on should judgments within a condition? Subjects were divided within each condition based on their answers to the on purpose question. There were no significant difference in answers to the should question across subjects who answered the on purpose question in the affirmative or negative (Upton’s $\chi^2<1, ps>.1$, two-tailed; see Figure below).
Figure 4.5: Percentage of subjects responding “yes” to the should question. Subjects are grouped by condition and their response to the intention to one question. There is a clear main effect of condition, suggesting that condition impacted answers to the should question independent of answers to the intention question.

Subjects were further subdivided based on whether they spontaneously mentioned an action the protagonist could have taken that would have led to a morally preferable outcome (a better alternative). Twenty-three subjects mentioned better alternatives, 14 in the Save Some Condition, 8 in the Better Alternative Available Condition, and 1 in the Save All Condition. Subject responses to the should, action rating, and on purpose questions were analyzed with “mention of better alternative” (“Mention”) as a categorical predictor in addition to Condition. A logistic regression was performed to determine the effect of Mention and Condition on answers to the should and on purpose questions. Neither model was significant (Should: $\chi^2(3) = 4.17, p = .24$; On Purpose: $\chi^2(3) = 4.0, p = .266$) and no predictor had an independent significant effect on outcome variables, nor
were there significant interactions. An ANOVA was performed to determine the effect of Mention and Condition on action ratings. The overall model was significant (F(5, 119) = 6.35, p<.0001), but neither predictor nor their interaction was significant. Condition was marginally significant (F(2, 119) = 2.19, p =.11).

Due to the low subject numbers mentioning better alternatives, post-hoc analysis was conducted by condition (the Save All Condition only had one mention of a better alternative, so analysis on that condition was omitted.)

In the Save Some Condition, 57.1% of subjects (8 out of 14) who mentioned a better alternative responded yes to the question “Should she have done that?” as compared to 84.8% (28 out of 33) of those in the Save Some Condition who did not mention a better alternative. In the Better Alternative Available Condition, 25.0% of subjects (2 out of 8) who mentioned a better alternative responded yes to the question “Should she have done that?” as compared to 28.1% (9 out of 32) of those in the Better Alternative Available Condition who did not mention a better alternative.

The difference in should judgments between the “mentioners” and the “non-mentioners” was significant in the Save Some Condition (Upton’s χ²(1, n = 47)= 4.12, p =.04, two-tailed) and not in the Better Alternative Available Condition (Upton’s χ²(1, n = 40)= .031, p =.86, two-tailed; Figure 5). No significant differences between mentioners and non-mentioners were found for either the action ratings or the on purpose question for either condition.
Figure 4.6: Subjects’ responses to the question, “Should she have done that?”

Figure 4.7: Subjects’ ratings of the action of the protagonist. Error bars shows standard error of the mean.
Figure 4.8: Subjects’ responses to the question, “Did she make the one kid happy/sad on purpose?”

**Experiment 2**

In Experiment 2, we tested adults on the same stimuli used with preschoolers in Experiment 1.

**Methods**

**Procedure.** Subjects were recruited through the Amazon Mechanical Turk Website and paid for their participation. Subjects received the same stimuli as preschoolers, but instead of a live experimenter, a recorded voice narrated the story. Subjects saw the animation with accompanying narration once and then answered a series of control questions about the story. There was no effect on the results if subjects were excluded
based on their answers to the controls, so all subjects were included in the analysis. (See Appendix D for list of control questions.)

Subjects were asked four questions, the first concerning the permissibility of the protagonist’s action (“Should she have done that?”), the second asking for a Likert scale rating of the agent’s action (“Please rate what Sally did on the following scale: Really good, A little good, Just OK, A little bad, Really bad.”), the third concerning the protagonist’s intention towards the bystander (“Did Sally make the one kid sad/happy on purpose?”), and the fourth concerning her intention towards the group of children (“Did Sally make all the kids on the right happy on purpose?”).

Subjects. One hundred eighty-six adults participated in the study. Forty-three subjects completed the Save Some Condition, 74 completed the Better Alternative Available Condition, and 69 completed the Save All Condition.

Results

In the Save Some Condition, 76.7% of subjects (33 out of 43) responded yes to the question “Should she have done that?” (binomial test, N=43, x=10, p=.0003, two-tailed). In the Better Alternative Available Condition, 23.0% of subjects (17 out of 74) said yes to the “should” question (binomial test, N=74, x=17, p<.0001, two-tailed). In the Save All Condition, 97.1% of subjects (67 out of 69) said yes to the “should” question (binomial test, N=69, x=2, p<.0001, two-tailed).

There was a significant difference between subjects’ answers to the should question across the three conditions (Pearson $\chi^2(1, n = 186) = 88.7, p <.0001$, two-tailed).
Subjects were significantly more likely to answer *yes* to the “should” question in the Save Some Condition than the Better Alternative Condition (Upton’s $\chi^2(1, n = 117)= 33.9$, $p <.0001$, two-tailed) and in the Save All Condition than the Better Alternative Condition (Upton’s $\chi^2(1, n = 143)= 80.4$, $p <.0001$, two-tailed).

Figure 4.9: Subjects’ responses to the question, “Should she have done that?”

Likert scale ratings were scored as follows: Really Bad = -2; A Little Bad = -1; Just OK = 0; A Little Good = 1; Really Good = 2. Subjects in the Save Some Condition rated the action as slightly good (M= .58; SE=.17). Subjects in the Better Alternative Available Condition rated the action as slightly bad (M= -.55; SE=.19). Subjects in the Save All condition rated the action as Really Good (M=1.88; SE=.06).

Analysis of variance revealed a significant difference between the conditions, $F(2, 186) = 123.1$, $p<.0001$. Ratings in the Better Alternative Condition were significantly worse than ratings in the Save Some Condition (independent-sample t-test, $t(115)=4.03$, ...
p<.0001, two-tailed) and significantly worse than ratings in the Save All Condition (independent-sample t-test, t(141)=11.9, p<.0001, two-tailed).

Figure 4.10: Subjects’ ratings of the action of the protagonist. Error bars shows standard error of the mean.

In the Save Some Condition, 27.9% of subjects (12 out of 43) responded yes to the question “Did she make that one kid happy/sad on purpose?” (binomial test, N=43, x=12, p=.001, two-tailed). In the Better Alternative Available Condition, 59.5% of subjects (44 out of 74) said yes to this question (binomial test, N=74, x=30, p=.065, two-tailed). In the Save All Condition, 84.1% of subjects (58 out of 69) said yes to the question “Did she make that one kid happy on purpose?” (binomial test, N=69, x=11, p<.0001, two-tailed).

There was a significant difference between subjects’ answers to the one on purpose question across the three conditions (Pearson $\chi^2(1, n = 186)= 35.4, p <.0001$, two-tailed). Subjects were significantly more likely to say that the protagonist made the one child sad on purpose in the Better Alternative Condition compared to the Save Some Condition (Upton’s $\chi^2(1, n = 117)= 10.76, p =.001$, two-tailed). Subjects were significantly more likely to say that the protagonist’s effect on the one child was intended
in the Better Alternative Condition than in Save All Condition (Upton’s $\chi^2(1, n = 143) = 10.49, p = .001$, two-tailed).

Figure 4.11: Subjects’ responses to the question, “Did she make the one kid happy/sad on purpose?”

In the Save Some Condition, 87.0% of subjects (60 out of 69) responded yes to the question “Did Sally make all the kids on the right happy on purpose?” (binomial test, N=43, x=9, p<.0001, two-tailed). In the Better Alternative Available Condition, 77.0% of subjects (57 out of 74) said yes to that question (binomial test, N=74, x=17, p<.0001, two-tailed). In the Save All Condition, 87.0% of subjects (60 out of 69) said yes (binomial test, N=69, x=9, p<.0001, two-tailed).

There were no significant differences between subjects’ answers to the three on purpose question across the three conditions (Pearson $\chi^2(1, n = 186) = 2.47, p = .29$, two-tailed).
Figure 4.12: Subjects’ responses to the question, “Did Sally make all the kids on the right happy on purpose?”

Is it possible that once intention is accounted for, there is no effect of condition on moral permissibility judgments (should question)? A logistic regression was performed to determine the effect of intention (to the one) and condition on answers to the should question. The model was significant ($\chi^2(3) = 42.6$, p<.0001) and, critically, each predictor had an independent significant effect on the outcome variable, as did their interaction (p<.0001).
Figure 4.13: Percentage of subjects responding “yes” to the *should* question. Subjects are grouped by condition and their response to the *intention to one* question. There is a clear main effect of condition, suggesting that condition impacted answers to the *should* question independent of answers to the *intention* question.

**Discussion**

The key finding in Experiments 1 and 2 is that the very same action can have a dramatically different moral status given a different class of choices available to the agent. This is preliminary evidence that the routine consideration of moral alternatives is a key component of moral cognition, in addition to the consideration of outcomes and intentions.

Is it possible that the considerations of alternatives do not play a separate role in moral cognition, but instead simply act to provide us with information about an actor’s intention, thereby playing an indirect role in moral judgment? In the Better Alternative Available condition, the protagonist has the option of putting up her gate in such a way that all the children would be protected, but instead she chooses to protect only some of the children. If subjects assume that the protagonist is a rational actor, what could possibly explain this behavior? One appealing explanation is that Sally *intends* for the squirrel to eat the lone bystander’s cookie (while also possibly intending the good effect of protecting the group of three children). The fact that Sally has the option of protecting everyone and does not, provides information about Sally’s harmful intention. In this way, is it possible that intention inference could entirely explain moral judgments?
Our results suggest that this is not the case. First, preschoolers do not judge the intention of the protagonist differently based on her choice situation, though they do make significantly different moral judgments based on choices available. Second, adults do make different intention inferences based on the availability of choices, though there is still an independent effect of the availability of choices on moral judgment, regardless of their intention inference.

In Experiment 1, children’s intention judgments did not differ across the conditions, nor did intention judgments predict answers to the *should* question within a condition, in contrast to the way that adults responded to the stimuli. Instead, preschoolers were at chance in judging the intention of the protagonist towards the bystander across all conditions. Why do preschoolers in our studies not seem to be making the same intention inferences that adults do? This is a somewhat surprising finding considering that even preverbal infants use the availability of alternatives to infer the preferences and intentions of agents (Woodward, 2013).

At least three explanations are possible. First, it is possible that the choices available in our experiment are too complex for preschoolers to conduct intention inference over them and subjects’ response to the *on purpose* question is at chance as a reflection of their confusion. This explanation is perhaps unlikely given that subjects process the choices sufficiently to use them in their moral judgments. Second, it is possible that preschoolers do in fact make the intention inferences that adults do, but that the prompt used in the study did not allow them to express their judgments adequately. This also seems unlikely given the success of the “on purpose” question in previous work (Leslie, Knobe, & Cohen, 2006). Third, it is possible that the information presented in
the story is not sufficient for preschoolers to draw conclusions about the protagonist’s intention towards the bystander. For example, they may represent the protagonist as having the general intention to protect the group of children or to protect as many children as possible and acknowledge (through their moral judgments) that she did not act optimally given her intention, though subjects’ may not represent her as having a specific intention towards the bystander. In other words, the protagonist’s intention towards the bystander goes unrepresented.

**Conclusion**

These studies suggest that when subjects are explicitly told about better alternatives an agent could have taken (or spontaneously think of them), they use this information in their moral evaluation of the agent. Future research should investigate 1) whether the agent’s awareness of a better alternative impacts a third party’s moral judgment and 2) which alternatives get considered and why.
V. Conclusion

Each of Chapters 2-4 in this dissertation looked at the relationship between moral rules and representations from a different perspective.

Chapter 2 highlighted the fact that a well-studied moral rule (harm as an unintended side-effect is permitted in certain cases) operates over a representation of the intention of the moral agent, though no proposal concerning how that representation is built from an impoverished stimulus (ie, lacking intention information) has been empirically tested. Put another way, the question of how the stimulus is converted to a mental representation that contains all the necessary information for the application of the moral rule has mostly been ignored – and entirely ignored from an experimental perspective. Indeed, the question of how we determine the intention of an agent in a world in which agents move around causing myriads of effects is a problem shackling all theories of intention inference. We proposed that this problem may become tractable if we look to domain-specific solutions. In particular, in the moral domain, we propose that we have a prior on agents intending the good effects and not the bad effects of their actions. Given this prior, a representation of the agent’s intention towards the good and bad effects of her action can be built, even when that explicit information is absent from the stimulus. This allows the moral rule (unintended harmful side-effects may be permitted) to be applied.

Chapter 3 used the strategy of looking at two abstract moral rules (the means principle and the implied consent principle) in order to understand how preschoolers are representing the moral world. The data suggested that preschoolers use the means principle in their moral judgments, treating harm committed as a means as morally worse than that same harm committed as a foreseen side-effect. That preschoolers are able to
use this rule hints at the fact that they represent morally charged effects as being imbedded in hierarchical action structures, composed of sub- and superordinate goals. After all, the difference between a means and a side-effect is that event’s place in the causal and intention structure of the action. An event is a means if it is causally necessary and intended (that is, a sub-goal); a side-effect is neither causally necessary nor intended.

Chapter 3 did not explicitly address the following important question: is it possible that preschoolers in Experiment 1 (which tested the means/side-effect distinction) did not represent the content of the goal of the agent (saving the five children), but only detected that harm was caused as a means or side-effect of some or other goal? It is possible that detecting the presence of harm as a means and harm as a side-effect (irrespective of the ultimate goal of the agent) would be sufficient for subjects to judge the means case impermissible and the side-effect case permissible. This would suggest that subjects are representing hierarchical action in a somewhat impoverished way, with action structure intact (there is a representation that some super-ordinate goal exists) but without the full content.

However, the results from Chapter 3, Experiments 3-4 (the implied consent studies) suggests that this is not the case. The implied consent stories involve harm committed as a means: the protagonist takes another child’s ball as a causally necessary and intended step in her plan to protect that child. Yet these cases are judged to be permissible. The difference, of course, between the implied consent cases and the cases of harm as a means in Experiment 1 is the content of the superordinate goal of the action. In the implied consent cases, the superordinate goal involves protecting the same person
that is being harmed, while the superordinate goal in the means case from Experiment 1 involves protecting five other individuals. The first thing to note is that the identity of the victim – whether or not the victim is also the beneficiary of the good effect – seems to make a moral difference. Second, the fact that we can detect this moral difference suggests that subjects seem to be representing the content of the superordinate goal of the agent, not merely that some superordinate goal exists.

Together, the studies in Chapter 3 showed how investigating moral rules can hint at the way that the moral world is represented. In a similar vein, Chapter 4 looked at how a moral rule (harm as a side-effect is permitted if there is no better alternative) hints at the way subjects represent the moral world. Chapter 4 argued that in addition to the well-studied features of intention and outcome, the presence of alternatives are represented as relevant to the moral permissibility of the agent’s action. Chapter 4 showed that the very same moral act can be permitted if it is the best alternative, but forbidden if there is a morally better alternative. It is an open question whether the agent must be aware of the alternatives for the subject’s moral judgment of her action to take those alternatives into account.

There is, as always, more work to be done on each of the issues opened here. At the very least, this has been the start of “discover[ing] the representations that appear and the rules operating on them” in the developing moral mind (Chomsky, 1980).
Appendix A

Pink Scale Training

Children were shown the following scale:

This is called the Pink Scale Game and in this game we show each other when things are good [point to stars], bad [point to x’s] or just ok [point to circle]. First let’s think of something good. Can you think of something good? [Wait for child to respond.] Is that really good [point to lots of stars] or just a little good [point to one star]? Then encourage child to offer a suggestion of something that is a little good or really good, whichever they haven’t already offered.

If child can’t think of something good at all (or can’t think of something really good or a little good) offer a suggestion such as “eating an apple” or “helping your teacher” or “playing outside.”

Repeat with “bad” and “just OK”.

Children are then told two stories accompanied by pictures:
This is a story about Billy and Johnny. In this story, Billy hits Johnny.

Should Billy have done that?

Can you show me on the pink scale? Was what Billy did: good, bad, or just OK?

This is a story about Sue and Anne. What is Anne holding? That’s right, a flower! In this story, Anne gives her flower to Sue.

Should Anne have done that?

Can you show me on the pink scale? Was what Anne did: good, bad, or just OK?

To be included in the study, children needed to get both answers correct for both a good and bad story. If children failed the bad story, they were given another bad story; if they failed the good story, they were given another good story (below):

This is a story about Billy and Johnny. In this story, Billy has a cookie and he gives it to Johnny.

Should Billy have done that?

Can you show me on the pink scale? Was what Billy did: good, bad, or just OK?

This is a story about Sue and Anne. What is Anne holding? That’s right, a flower! In this story, Sue takes Anne’s flower and she breaks it.

Should Anne have done that?
Can you show me on the pink scale? Was what Anne did: good, bad, or just OK?
Appendix B

Stimuli for Chapter II Studies

Stimuli for Experiment 1: Adult Subjects

Neutral Condition

Hank is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Hank sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Hank is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the five men. There is a man on the side track. Hank can throw the switch, killing him; or he can refrain from doing this, letting the five die. Hank throws the switch.

Good Condition

Joe is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Joe sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Joe is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the five men. There is a man on the side track. Joe can throw the switch, killing him; or he can refrain from doing this, letting the five men die. Joe then recognizes that the five men are
people who he deeply cares about. Joe thinks to himself, “This is my chance to save those people.” Joe throws the switch.

**Bad Condition**

Mark is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Mark sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Mark is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the five men. There is a man on the side track. Mark can throw the switch, killing him; or he can refrain from doing this, letting the five men die. Mark then recognizes that the man on the side track is someone who he hates with a passion. Mark thinks to himself, “This is my chance to kill that bastard.” Mark throws the switch.

**Test Questions for Experiment 1:**

Is it morally permissible for Hank/Joe/Mark to throw the switch? (Options: Yes/No)

Hank/Joe/Mark threw the switch in order to kill the man. (Options: True/False)

**Stimuli for Experiment 2: Preschool Subjects**

These stories were accompanied by animations shown to the subjects.

**Neutral Condition**

This is a story about Sally. And Sally is playing in the park. And there are some other kids in this story too. There is one kid over here. And there are lots of kids over here.
See this one kid? This is a new kid. She has never been to the park before. Sally has never met her.

Does Sally know this kid?

*If correct, say “That’s right, Sally does not know this kid.”*

*If incorrect, say “Now listen carefully” and repeat story*

See all these kids? These are new kids. They have never been to the park before. Sally has never met them.

Does Sally know these kids?

*If correct, say, “That’s right, Sally does not know these kids.”*

*If incorrect, say “Now listen carefully” and repeat story*

Today, all the kids in the park are eating cookies. They are all eating cookies! But uh oh, here comes a mean sneaky squirrel who likes to eat other people’s food.

Can you tell where he wants to go?

*If correct, say, “That’s right! The squirrel is going to go eat all those kids’ cookies!”*

*If incorrect, ask which way the squirrel is looking*

*If still incorrect, say, “He is going to eat these kids’ cookies over here.” Point to 5.*

And if the squirrel eats their cookies, how will these kids feel?

*If they give any negative affect emotion (sad, bad, mad) say, “That’s right, they’ll feel sad.”*

*If incorrect or no answer say, “They’ll be sad if the squirrel eats their cookies.”*
Well, Sally knows what the squirrel is going to do. Sally knows that the squirrel is going to go eat those kid’s cookies and make them sad. So, let’s see what she does! Sally has a gate with her, and she decides to put the gate right there. She knows that now the squirrel can’t reach all these kids’ cookie. So he is going to go over here and eat this kid’s cookie instead. So this kid is sad because he doesn’t get to eat his own cookie. But these kids aren’t sad because they get to eat their own cookies.

Let’s watch that again. [Replay video from the start.]

*If subjects don’t remember, help them. “Where is the squirrel looking? Whose cookies did he want to eat?”*

*If correct response, say “That’s right.”*

How were these kids going to feel?

*If subjects don’t remember, help them.*

*If correct response, say “That’s right.”*

**Exclusion Criteria:**

What did Sally do?

What did the squirrel do?

How did that kid feel?

Were those kids sad?

**Test questions:**
Ok, that’s the end of the story. But, I’m wondering about something. I’m wondering about Sally and what she did. See this sad kid? [Point to the one.] Did Sally make this kid sad on purpose?

In this story Sally used her gate. SHOULD she have done that?

Can you show me on the Pink Scale? Was what Sally did good, bad, or just OK?

**Good Condition**

This is a story about Sally. And Sally is playing in the park. And there are some other kids in this story too. There is one kid over here. And there are lots of kids over here.

See this one kid? This is a new kid. She has never been to the park before. Sally has never met her.

Does Sally know this kid?

*If correct, say “That’s right, Sally does not know this kid.”*

*If incorrect, say “Now listen carefully” and repeat story*

See all these kids? Sally really likes these kids. These kids are Sally’s friends. Sally likes these kids a lot.

Does Sally like these kids?

*If correct, say, “That’s right, Sally likes these kids.”*

*If incorrect, say “Now listen carefully” and repeat story*
Today, all the kids in the park are eating cookies. They are all eating cookies! But uh oh, here comes a mean sneaky squirrel who likes to eat other people’s food.

Can you tell where he wants to go?

*If correct, say, “That’s right! The squirrel is going to go eat all those kids’ cookies!”*

*If incorrect, ask which way the squirrel is looking*

*If still incorrect, say, “He is going to eat these kids’ cookies over here.” Point to 5.*

And if the squirrel eats their cookies, how will these kids feel?

*If they give any negative affect emotion (sad, bad, mad) say, “That’s right, they’ll feel sad.”*

*If incorrect or no answer say, “They’ll be sad if the squirrel eats their cookies.”*

Well, Sally knows what the squirrel is going to do. Sally knows that the squirrel is going to go eat those kid’s cookies and make them sad. But remember, Sally likes these kids. Sally doesn’t want the squirrel to eat these kids’ cookies. Sally doesn’t want these kids to be sad.

So, let’s see what she does! Sally has a gate with her, and she decides to put the gate right there. She knows that now the squirrel can’t reach all these kids’ cookie. So he is going to go over here and eat this kid’s cookie instead. So this kid is sad because he doesn’t get to eat his own cookie. But these kids aren’t sad because they get to eat their own cookies.

Let’s watch that again. [Replay video.]

At the beginning, where was the squirrel going to go?
If subjects don’t remember, help them. “Where is the squirrel looking? Whose cookies did he want to eat?” If correct response, say “That’s right.” How were these kids going to feel?

If subjects don’t remember, help them. If correct response, say “That’s right.”

Exclusion Criteria:

Does Sally like these kids? (the five)

How does Sally want to make these kids feel?

What did Sally do?

What did the squirrel do?

How did that kid feel?

Were those kids sad?

Test Questions:

Ok, that’s the end of the story. But, I’m wondering about something. I’m wondering about Sally and what she did. See this sad kid? [Point to the one.] Did Sally make this kid sad on purpose?

In this story Sally used her gate. SHOULD she have done that?

Can you show me on the Pink Scale? Was what Sally did good, bad, or just OK?

Bad Condition

This is a story about Sally. And Sally is playing in the park. And there are some other kids in this story too. There is one kid over here. And there are lots of kids over here.
See this one kid? Sally doesn’t like this kid. Sally doesn’t like this kid one bit. They are not friends.

Does Sally like this kid?

*If correct, say “That’s right, Sally does not like this kid”.*

*If incorrect, say “Now listen carefully” and repeat story*

See all these kids? These are new kids. They have never been to the park before. Sally has never met them.

Does Sally know these kids?

*If correct, say, “That’s right, Sally does not know these kids.”*

*If incorrect, say “Now listen carefully” and repeat story*

Today, all the kids in the park are eating cookies. They are all eating cookies! But uh oh, here comes a mean sneaky squirrel who likes to eat other people’s food.

Can you tell where he wants to go?

*If correct, say, “That’s right! The squirrel is going to eat all those kids’ cookies!”*

*If incorrect, ask which way the squirrel is looking*

*If still incorrect, say, “He is going to eat these kids’ cookies over here.” Point to 5.*

And if the squirrel eats their cookies, how will these kids feel?

*If they give any negative affect emotion (sad, bad, mad) say, “That’s right, they’ll feel sad.”*

*If incorrect or no answer say, “They’ll be sad if the squirrel eats their cookies.”*
Well, Sally knows what the squirrel is going to do. Sally knows that the squirrel is going to go eat those kid’s cookies and make them sad. But remember, Sally doesn’t like this kid. Sally wants the squirrel to eat this kid’s cookie. Sally wants this kid to be sad.

So, let’s see what she does! Sally has a gate with her, and she decides to put the gate right there. She knows that now the squirrel can’t reach all these kids’ cookie. So he is going to go over here and eat this kid’s cookie instead. So this kid is sad because he doesn’t get to eat his own cookie. But these kids aren’t sad because they get to eat their own cookies.

Let’s watch that again. [Replay video.]

At the beginning, where was the squirrel going to go?

*If subjects don’t remember, help them. “Where is the squirrel looking? Whose cookies did he want to eat?”*

*If correct response, say “That’s right.”*

How were these kids going to feel?

*If subjects don’t remember, help them.

*If correct response, say “That’s right.”*

**Exclusion Criteria:**

Does Sally like this kid? (the one)

How does Sally want to make this kid feel?

What did Sally do?

What did the squirrel do?

How did that kid feel?
Were those kids sad?

Test Questions:

Ok, that’s the end of the story. But, I’m wondering about something. I’m wondering about Sally and what she did. See this sad kid? [Point to the one.] Did Sally make this kid sad on purpose?

In this story Sally used her gate. SHOULD she have done that?

Can you show me on the Pink Scale? Was what Sally did good, bad, or just OK?
Appendix C
Stimuli for Chapter III Studies

Experiment 1: Materials and Methods

Side-Effect Condition

This is a story about Katy. And Katy is playing in the park. There are some other kids in this story too. There is one kid over here and lots of kids over here. This girl over here as a gate with her. She just carries around her gate with her just like that. Today, all these kids are eating cookies! These kids all have cookies. But ut oh, here comes a mean, sneaky squirrel. He likes to eat other people’s food. That squirrel is coming over here and -- what do you think he’s going to do? That’s right, he wants to come over here and eat all these kids cookies. Not this kid [point to the one], just these kids [the five]. And how are these kids going to feel if the squirrel eats their cookies? That’s right, they’ll feel sad. So Katy knows what the squirrel is going to do. Katy knows that the squirrel is going to go eat those kids cookies and make them feel sad. So let’s see what she does!

Katy waves to this girl over here. Katy waves to her and says, “Hey, bring your gate over here!” So that girl brings her gate over. Katy knows that when the girl gets here, that the squirrel will eat her cookie. Did Katy know that that would happen? [Correct if subject gets it wrong.] And Katy also knows that when the girl gets to her, that her gate will block the squirrel. And the squirrel cannot get passed. Did Katy know that that would happen? [Correct if subject gets it wrong.] So that girl is sad because she doesn’t get to eat her own cookie. But these kids are not sad, because they get to eat their own cookies.

Ok, let’s watch that again from the beginning. [Replay the animation.]
In the beginning, do you remember what this girl has with her?

Where was the squirrel going to go?

**Control questions (subjects were excluded for incorrect answers):**

How were those kids going to feel?

What did Katy do?

Where did the girl go?

What did the squirrel do?

Did Katy know that that would happen?

How did that girl feel?

Did the squirrel get passed the gate?

Did Katy know that that would happen?

Did these kids feel sad?

**Test questions:**

So that’s the end of the story, but I’m wondering about something. I’m wondering about Katy and what she did. In this story, Katy waved to the girl and told her to come over. Should she have done that? Can you show me on the pink scale. Was what Katy did, good, bad, or just OK?

**Means Condition**

This is a story about Katy. And Katy is playing in the park. There are some other kids in this story too. There is one kid over here and lots of kids over here. Today, all these kids are eating cookies! These kids all have cookies. But ut oh, here comes a mean, sneaky squirrel. He likes to eat other people’s food. That squirrel is coming over here and -- what do you think he’s going to do? That’s right, he wants to come over here and eat all these kids cookies. Not this kid [point to the one], just these kids [the five]. And how are these kids going to feel if the squirrel eats their cookies? That’s right, they’ll feel sad. So Katy knows what the squirrel is going to do. Katy
knows that the squirrel is going to go eat those kids cookies and make them feel sad. So let’s see what she does!

Katy waves to this girl over here. Katy waves to her and says, “Hey, come over here!” And Katy knows that when the girl gets to here, that she is going to take her cookie and feed it to the squirrel. And the squirrel will eat her cookie. So that girl is sad because she doesn’t get to eat her own cookie. But these kids are not sad, because they get to eat their own cookies.

Ok, let’s watch that again from the beginning. [Replay the animation.]

In the beginning, do you remember where the squirrel was going to go?

How were those kids going to feel?

Control questions (subjects were excluded for incorrect answers):

What did Katy do?
Where did the girl go?
What did the squirrel do?
How did that girl feel?
Did these kids feel sad?

Test questions:

So that’s the end of the story, but I’m wondering about something. I’m wondering about Katy and what she did. In this story, Katy took this girl’s cookie and fed it to the squirrel. Should she have done that? Can you show me on the pink scale. Was what Katy did, good, bad, or just OK?

Experiment 1a: Materials and Methods
This is a story about Katy. And Katy is playing in the park. There are some other kids in this story too. There is one kid over here and lots of kids over here. Today, all these kids are eating cookies! These kids all have cookies. But ut oh, here comes a mean, sneaky squirrel. He likes to eat other people’s food. That squirrel is coming over here and -- what do you think he’s going to do? That’s right, he wants to come over here and eat all these kids cookies. Not this kid [point to the one], just these kids [the five]. And how are these kids going to feel if the squirrel eats their cookies? That’s right, they’ll feel sad. So Katy knows what the squirrel is going to do. Katy knows that the squirrel is going to go eat those kids cookies and make them feel sad. So let’s see what she does!

Katy waves to this girl over here. Katy waves to her and says, “Hey, come over here!” So that girl comes over. And Katy knows that when the girl gets to here, that she will get in the way of the squirrel. And the squirrel will eat her cookie. So that girl is sad because she doesn’t get to eat her own cookie. But these kids are not sad, because they get to eat their own cookies.

Ok, let’s watch that again from the beginning. [Replay the animation.] In the beginning, do you remember where the squirrel was going to go? How were those kids going to feel?

Control questions (subjects were excluded for incorrect answers):

What did Katy do?
Where did the girl go?
What did the squirrel do?
How did that girl feel?
Did these kids feel sad?
Test questions:
So that’s the end of the story, but I’m wondering about something. I’m wondering about Katy and what she did. In this story, Katy waved to the girl and told her to come over. Should she have done that? Can you show me on the pink scale. Was what Katy did, good, bad, or just OK?

Experiment 2: Materials and Methods

Side-Effect Condition
This is a story about Jimmy. And Jimmy is going for a walk in the forest. Woooo – there he goes into the forest! When Jimmy gets to the forest, what does he see? That’s right, apples! There is one apple over here and lots of apple over here. But ut oh, here comes a really rumbly rock that’s rolling down this hill. That rock is going to come over here and squish all of these apples. Not this one [point to the one], just these apples [the five]. Jimmy knows what the rock is going to do. Jimmy knows that the rock is going to come over here and squish all of these apples. So let’s see what he does!

Well, Jimmy goes and gets his ramp and he puts it right here. So now when the rock comes down, it rolls down the ramp. And squishes this apple [the one]. And these apples are not squished. And then Jimmy goes away.

Ok, let’s watch that again from the beginning. [Replay the animation.]
In the beginning, what did Jimmy see?

Control questions (subjects were excluded for incorrect answers):
Where is the rock going to go?
What was going to happen to the apples?
What did Jimmy do?
Where did the rock go?
What happened to that apple?
Were those apples squished?

Test questions:
So that’s the end of the story, but I’m wondering about something. I’m wondering about Jimmy and what he did. In this story, Jimmy used his ramp. Should he have done that? Can you show me on the pink scale. Was what Jimmy did, good, bad, or just OK? See this squished apple?
Did Jimmy squish that apple on purpose?

Means Condition
This is a story about Jimmy. And Jimmy is going for a walk in the forest. Woooo – there he goes into the forest! When Jimmy gets to the forest, what does he see? That’s right, apples! There is one apple over here and lots of apple over here. But ut oh, here comes a really rumbly rock that’s rolling down this hill. That rock is going to come over here and squish all of these apples. Not this one [point to the one], just these apples [the five]. Jimmy knows what the rock is going to do. Jimmy knows that the rock is going to come over here and squish all of these apples. So let’s see what he does!

Well, Jimmy picks up this apple over here and when the rock comes down, he throws the apple at the rock. And this apple [the one] is squished. And these apples are not squished. And then Jimmy goes away.
Ok, let’s watch that again from the beginning. [Replay the animation.]

In the beginning, what did Jimmy see?
Control questions (subjects were excluded for incorrect answers):

Where is the rock going to go?
What was going to happen to the apples?
What did Jimmy do?
Where did the rock go?
What happened to that apple?
Were those apples squished?

Test questions:

So that’s the end of the story, but I’m wondering about something. I’m wondering about Jimmy
and what he did. In this story, Jimmy threw the apple at the rock. Should he have done
that? Can you show me on the pink scale. Was what Jimmy did, good, bad, or just OK? See this
squished apple? Did Jimmy squish that apple on purpose?

Experiment 3: Materials and Methods

Implied Consent Condition

This is a story about Jane. And Jane is going for a walk on the beach. She sees this girl who has
all of these beach balls. Look how many beach balls she has! The girl puts down the beach balls
so she can play with them. But uh-oh, here comes a mean, scary crab with big, scary claws. That
crab likes to break toys. And that crab is coming after those beach balls. The crab is going to
come over here and break all the beach balls. How will the girl feel if the crab breaks her beach
balls? That’s right, she’ll feel sad. So let’s see what happens. Well the crab comes over and
breaks the first ball and then the second ball and then all of the other beach balls. And this girl is
sad because she doesn’t get to play with her beach balls today.
The next day, Jane is going for a walk on the beach again. She sees this girl again who has all of these beach balls. The girl puts down the beach balls so she can play with them. But uh-oh, here comes that mean, scary crab again with big, scary claws. And that crab is coming after those beach balls. How will the girl feel if the crab breaks her beach balls? That’s right, she’ll feel sad. So let’s see what happens. Well the crab comes over and breaks the first ball and then the second ball and then all of the other beach balls. And this girl is sad because she doesn’t get to play with her beach balls today.

The next day, Jane is going for a walk on the beach again. She sees this girl again who has all of these beach balls. The girl puts down the beach balls so she can play with them. But uh-oh, here comes that mean, scary crab again with big, scary claws. And that crab is coming after those beach balls. Jane knows what the crab is going to do. She knows the crab is going to come and break the beach balls and make this girl sad. So, let’s see what she does.

Well, Jane comes over and takes the girl’s ball. And she puts the ball over here. And Jane knows that now the crab will chase the ball away. So the crab is gone. And he is never coming back. So the girl can go home with the rest of her beach balls.

Alright, now let’s watch that story one more time. [Replay last scene of story.]

Control questions (subjects were excluded for incorrect answers):

Do you remember in the beginning what the crab was going to do?

And if he did that, how was the girl going to feel?

What did Jane do?

And where did the crab go?

And is the crab coming back?

So where did the girl go?
Test questions:

In this story, Jane took the ball and she put it over there. Should Jane have done that? [Ask this question with the picture on the screen that shows Jane throwing the ball away.]

Can you show me on the Pink Scale? Was what she did good, bad, or just okay?

Do you remember how the crab was going to break the beach balls? What would be better for this girl? Would it be better for the crab to break the balls or for Jane to take the ball?

Baseline Condition

This is a story about Jane. And Jane is going for a walk on the beach. She sees this girl who has all of these beach balls. Look how many beach balls she has! The girl puts down the beach balls so she can play with them. Jane comes over and takes the girl’s ball. And she puts the ball over here. And the girl goes home with the rest of her beach balls.

[Tell the whole story again.]

Test questions:

In this story, Jane took the ball and she put it over there. Should Jane have done that? [Ask this question with the picture on the screen that shows Jane throwing the ball away.]

Can you show me on the Pink Scale? Was what she did good, bad, or just okay?

Experiment 4: Materials and Methods

Implied Consent Condition

This is a story about Jane. And Jane is going for a walk in the park. She sees this girl playing with
a ball. But uh-oh, here comes a mean, scary dog with big, scary teeth. That dog is coming after this girl. The dog is going to come over and bite the girl and hurt her and that girl is going to cry and cry. How will the girl feel? That’s right, she’ll feel sad. She’ll be hurt and she’ll cry and cry. Jane sees the dog and knows what the dog is going to do. She knows the dog will come over and hurt the girl and make her cry. Let’s see what she does.

Well, Jane comes up and takes the girl’s ball. And she puts the ball over here. And Jane knows that now the dog will chase after the ball. So the dog is gone. And he is never coming back. And the girl is safe. So the girl can go home.

Alright, now let’s watch that story one more time. [Replay animation.]

Control questions (subjects were excluded for incorrect answers):

Do you remember in the beginning what the dog was going to do?

And if he did that, how was the girl going to feel?

What did Jane do?

Where did the dog go?

And is the dog coming back?

So where did the girl go?

Test questions:

[Scroll back to the point in the story where the dog was about to attack the girl.] Do you remember how the dog was going to bite the girl? Would it be better for the dog to bite the girl or for Jane to take the ball?

In this story, Jane took the ball and she put it over there. Should Jane have done that? [Ask this question with the picture on the screen that shows Jane throwing the ball away.]

Can you show me on the Pink Scale? Was what she did good, bad, or just okay?
Baseline Condition

This is a story about Jane. And Jane is going for a walk in the park. She sees this girl playing with a ball. Well, Jane comes up and takes the girl’s ball. And she puts the ball over here. And the girl goes home.

[Tell the whole story again.]

Test questions:

In this story, Jane took the ball and she put it over there. Should Jane have done that? [Ask this question with the picture on the screen that shows Jane throwing the ball away.]

Can you show me on the Pink Scale? Was what she did good, bad, or just okay?
Appendix D

Stimuli for Chapter IV Studies

Experiment 1 (Preschoolers):

Save Some Condition
Subjects were shown the following video as the experimenter narrated the story using the script below. The experimenter advanced the narration at the pace appropriate to the child’s responses.

https://www.youtube.com/watch?feature=player_embedded&v=Knqrv3ZxbvI

This is a story about Sally and Sally is playing in the park. There are some other kids in this story too. There is one kid over here. What is that kid holding? That’s right, she’s holding a cookie! This kid is getting ready to eat her cookie. There are some more kids in this story, too. There are lots of kids over here. They are also getting ready to eat their cookies. But ut oh! Here comes a mean, sneaky squirrel who likes to eat other people’s food. Can you tell where he wants to go? That’s right, he’s going to go over here and eat all those kids’ cookies. And how will those kids feel if the squirrel eats their cookies? That’s right, they’ll feel sad. Well, Sally knows what the squirrel is going to do. Sally knows that the squirrel is going to go and eat those kids’ cookies and make them feel sad.

So what should she do? Hmmm.

Well, Sally has a gate with her.

And Sally could put her gate up right here [position A]. And if she puts it up here, then the squirrel won’t be able to get to all those kids’ cookies. So the squirrel will have to go over there and eat that one kid’s cookie instead. And that one kid will be sad.

So, let’s see what she does. Well, Sally puts the gate up right here [position A]. So, now the squirrel can’t get to all these kids’ cookies. So the squirrel has to come over here and eat this one kid’s cookie. So this one kid is sad, because she doesn’t get to eat her own cookie. But these kids are not sad, because they get to eat their own cookies.

Ok, let’s watch that again. [Play story again from the beginning.]

In the beginning, do you remember where the squirrel was going to go?
And how were those kids going to feel?
[ If wrong, help again. ]

Control questions: [Subjects excluded for giving incorrect answers.]
And if Sally put up her gate here, whose cookies was the squirrel going to eat?
[ If unclear, help them. Rephrase question, provide options, such as “who would get to eat their cookie?” allowing child to point. If wrong, do not correct. ]
So what did Sally do?
[ Help out if needed. Let them point to where Sally went, ask if she put something there. ]
What did the squirrel do?
[ If unclear, help them. Provide options, such as “can you show me where he went?”, allowing child to point. If wrong, do not correct. ]
And how did this kid feel?
[ If unclear, help them. Provide options, such as “was she happy or sad?” If wrong, do not correct. ]
How did these kids feel?
[ If unclear, help them. Provide options, such as “were they happy or sad?” If wrong, do not correct. ]

Test questions:
I’m wondering about what Sally did. In this story, Sally used her gate. Should she have done that?
Can you show me on the pink scale? Was what Sally did, good, bad, or just OK?
One last question. See this sad girl. Did Sally make this girl sad on purpose?

Better Alternative and Save All Conditions

Subjects were shown the following video as the experimenter narrated the story using the script below. The experimenter advanced the narration at the pace appropriate to the child’s responses.

Better alternative condition:
https://www.youtube.com/watch?feature=player_embedded&v=qdvYyLV8FJQ

Save All condition:
https://www.youtube.com/watch?feature=player_embedded&v=--gh3WSNiDA
This is a story about Sally and Sally is playing in the park. There are some other kids in this story too. There is one kid over here. What is that kid holding? That’s right, she’s holding a cookie! This kid is getting ready to eat her cookie. There are some more kids in this story, too. There are lots of kids over here. They are also getting ready to eat their cookies. But ut oh! Here comes a mean, sneaky squirrel who likes to eat other people’s food. Can you tell where he wants to go? That’s right, he’s going to go over here and eat all those kids’ cookies. And how will those kids feel if the squirrel eats their cookies? That’s right, they’ll feel sad. Well, Sally knows what the squirrel is going to do. Sally knows that the squirrel is going to go and eat those kids’ cookies and make them feel sad.

So what should she do? Hmmmm.

Well, Sally has a gate with her.

And Sally could put up her gate **up** right here [position A]. And if she puts it **up** here, then the squirrel won’t be able to get to all those kids’ cookies. So the squirrel will have to go over there and eat that one kid’s cookie instead. And that one kid will be sad.

Or she could put up her gate **down** here [position B]. And if she puts it **down** here, then the squirrel won’t be able to get to anybody’s cookies. And no one will be sad.

**Better Alternative Condition:**
So, let’s see what she does. Well, Sally puts the gate **up** right here [position A]. So, now the squirrel can’t get to all these kids’ cookies. So the squirrel has to come over here and eat this one kid’s cookie. So this one kid is sad, because she doesn’t get to eat her own cookie. But these kids are not sad, because they get to eat their own cookies.

**Save All Condition:**
So, let’s see what she does. Well, Sally puts the gate **down** right here [position B]. So, now the squirrel can’t get to anybody’s cookies. So the squirrel just goes away. So this one kid is not sad, because she gets to eat her own cookie. And these kids are not sad, because they get to eat their own cookies, too.

Ok, let’s watch that again. [Replay video.]

In the beginning, do you remember where the squirrel was going to go?
[ If wrong, help them. Do you remember, in the very beginning...? ]
And how were those kids going to feel?
[ If wrong, help again. ]

**Control Questions:** [Subjects excluded for giving incorrect answers.]
And if Sally put her gate up here, whose cookies was the squirrel going to eat?
[ If unclear, help them. Rephrase question, provide options, such as “who would get to eat their cookie?”, allowing child to point. If wrong, do not correct. ]

And if Sally put her gate down here, whose cookies was the squirrel going to eat?
[ If unclear, help again. If wrong, do not correct. ]

So what did Sally do?
[ Help out if needed. Let them point to where Sally went, ask if she put something there. ]

What did the squirrel do?
[ If unclear, help them. Provide options, such as “can you show me where he went?”, allowing child to point. If wrong, do not correct. ]

And how did this kid feel?
[ If unclear, help them. Provide options, such as “was she happy or sad?” If wrong, do not correct. ]

How did these kids feel?
[ If unclear, help them. Provide options, such as “were they happy or sad?” If wrong, do not correct. ]

**Test Questions:**
Ok, that’s the end of the story. But, I’m wondering about something. I’m wondering about Sally and what she did. In this story, Sally used her gate. [Point to the orientation of the gate.]
Should she have done that?
Can you show me on the pink scale? Was what Sally did, good, bad, or just OK?
One last question. See this sad girl. Did Sally make this girl sad on purpose?
Experiment 2 ( Adults):

Subjects watched the same video as the children. Instead of a live experimenter, a recorded voice (reading the same script as for the children) narrated the story.

Thank you for taking the time to complete this survey. For our research, we are interested in knowing certain facts about you. Specifically, we are interested in whether you read directions carefully. So, in order to show us that you have read the instructions, please ignore this first question about attending university (don’t answer it). Also, please copy and paste the words “I have read the instructions” (without the quotation marks) in the box labeled "Any comments or questions?” Thank you very much. Have you attended university?

- No, I did not attend
- Yes, and I graduated
- Yes, but I did not graduate

Any comments or questions?

Save Some Condition

Subjects watched the following video:

https://www.youtube.com/watch?feature=player_embedded&v=Knnqv3ZxbvI
The following questions will be about the story you just watched. Please refer to the accompanying pictures before each question.

In the very beginning of the story, where was the squirrel going to go?

- To the right, to eat all those kids' cookies.
- To the left, to eat that one kid's cookie.

And how were they going to feel?

- Happy
- Sad
If Sally put her gate up like this, whose cookies was the squirrel going to eat? (See picture above)

- The one kid on the left
- All the kids on the right

What did Sally do?

- She put the gate up:

- She did nothing:
Where did the squirrel go?

- To the one kid on the left
- To all the kids on the right

How did the one kid on the left feel?

- Happy
- Sad
And how did all the kids on the right feel?

- Happy
- Sad

In this story, Sally used her gate (see picture above). Should she have done that?

- Yes
- No

Please rate what Sally did on the following scale:

- Really good
- A little good
- Just okay
- A little bad
- Really bad
Did Sally make the one kid sad on purpose?

- Yes
- No

Did Sally make all the kids on the right happy on purpose?

- Yes
- No

**Better Alternative and Save All conditions:**
Subjects watched the follow video:

Better alternative condition:
https://www.youtube.com/watch?feature=player_embedded&v=qdvYyLV8FJQ

Save All condition:
https://www.youtube.com/watch?feature=player_embedded&v=--gh3WSNiDA

In the very beginning of the story, where was the squirrel going to go?

- To the right, to eat all those kids' cookies.
- To the left, to eat that one kid's cookie.
And how were they going to feel?

- Happy
- Sad

If Sally put her gate up like this, whose cookies was the squirrel going to eat? (See picture above)

[Order of gate presentation was counterbalanced.]

- The one kid on the left
- All the kids on the right
- No one
If Sally put her gate down like this, whose cookies was the squirrel going to eat? (See picture above)

[Order of gate presentation was counterbalanced.]

☐ The one kid on the left
☐ All the kids on the right
☐ No one
What did Sally do?

- She put the gate up
- She put the gate down
- She did nothing
Better Alternative Condition continues like this:

Where did the squirrel go?

- To the one kid on the left
- To all the kids on the right
- Away

How did the one kid on the left feel?

- Happy
- Sad
And how did all the kids on the right feel?

- Happy
- Sad

In this story, Sally used her gate (see picture above). Should she have done that?

- Yes
- No
Please rate what Sally did on the following scale:

- Really good
- A little good
- Just okay
- A little bad
- Really bad

Did Sally make the one kid sad on purpose?

- Yes
- No

Did Sally make all the kids on the right happy on purpose?

- Yes
- No
Save All Condition continues like this:

Where did the squirrel go?
- To the one kid on the left
- To all the kids on the right
- Away

How did the one kid on the left feel?
- Happy
- Sad
And how did all the kids on the right feel?

- Happy
- Sad

In this story, Sally used her gate (see picture above). Should she have done that?

- Yes
- No

Please rate what Sally did on the following scale:

- Really good
- A little good
- Just okay
- A little bad
- Really bad
Did Sally make the one kid sad on purpose?

☑ Yes
☒ No

Did Sally make all the kids on the right happy on purpose?

☑ Yes
☒ No
References


Dwyer, S. 2004: How good is the linguistic analogy. The innate mind, 2, 237-256.


Phillips, A. T., Wellman, H. M., & Spelke, E. S. (2002). Infants' ability to connect gaze and emotional expression to intentional action. *Cognition, 85*(1), 53-78.


