American Psychological Association

Journal of Experimental Psychology: General

© 2018 American Psychological Association
0096-3445/18/$12.00

2018, Vol. 147, No. 11, 1728–1747
http://dx.doi.org/10.1037/xge0000459

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

Sydney Levine
Rutgers University

John Mikhail
Georgetown University

Alan M. Leslie
Rutgers University

The presumption of innocence is not only a bedrock principle of American law, but also a fundamental human right. The psychological underpinnings of this presumption, however, are not well understood. To make progress, one important task is to explain how adults and children infer the goals and intentional structure of complex actions, especially when a single action has more than one salient effect. Many theories of moral judgment have either ignored this intention inference problem or have simply assumed a particular solution without empirical support. We propose that this problem may be solved by appealing to domain-specific prior knowledge that is either built-up over the probability of prior intentions or built-in as part of core cognition. We further propose a specific solution to this problem in the moral domain: a good intention prior, which entails a rebuttable presumption that if an action has both good and bad effects, the actor intends the good effects and not the bad effects. Finally, in a series of novel experiments we provide the first empirical support – from both adults and preschool children – for the existence of this good intention prior.

Keywords: moral development, intention inference, good intention prior, presumption of innocence, theory of mind

Supplemental materials: http://dx.doi.org/10.1037/xge0000459.supp

Agents move through the world constantly starting causal sequences of events that can be parsed in infinite ways. For example, you raise your hand signaling that you want to answer a question, but at the same time catch the eye of a student sitting behind you, 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 you intended. Yet the bare evidence supports inferences that any of them may have been your goal. This gives rise to a problem: How do individuals, from a very young age, reliably determine the goals of agents, given the vast number of parallel effects every action causes?

Some of the effects caused by an action seem immediately to 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 for now how salience gets attached to certain effects and not others. We instead focus on the problem of how individuals infer the intended goal of an action when more than one salient effect occurs.

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

In this article, we first review several theories of the development of intention inference, pointing out how each falls short of

Sydney Levine, Department of Psychology and Center for Cognitive Science, Rutgers University; John Mikhail, Georgetown University Law Center, Georgetown University; Alan M. Leslie, Department of Psychology and Center for Cognitive Science, Rutgers University.

Previous versions of this research (including the data and interpretations of the data) were presented at the Harvard Program on Psychiatry and the Law (2014), the meeting of the Society of Philosophy and Psychology (2014), and the Cognitive Development Society (2013).

We thank Fiery Cushman for valuable feedback. Parts of this research was supported by National Science Foundation Grant BCS-0922184 to Alan M. Leslie and the Department of Defense's National Defense Science & Engineering Fellowship to Sydney Levine. All data and stimulus materials are available at https://github.com/sydneylevine/good-intention-prior.

Correspondence concerning this article should be addressed to Sydney Levine, who is now at Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139. E-mail: smlevine@mit.edu

explaining how individuals can infer the intention of novel actions with more than one salient effect. We then propose a solution to this problem that may be particular to the moral domain, a good intention prior or what might be called a "presumption of innocence" (Mikhail, 2007, 2011). 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.

Goal Inference Theories

To begin, it is important to differentiate between two kinds of actions that might be labeled "intentional" (Premack, 1990; see also Searle, 1983). First, someone can act intentionally by acting voluntarily or nonaccidentally. Simply observing a white dot on a black background changing direction and speed is perceived as an intentional action performed by an animate agent (Tremoulet & Feldman, 2000; see Scholl & Tremoulet, 2000, for a review). However, perceiving this sort of "mechanical agency" (Leslie, 1994) does not necessarily require representing that the agent is acting toward a particular goal. On the other hand, 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 here is on theories that describe how individuals infer the latter sort of intention, in particular how they infer the goal of an action plan.

There is now a significant body of evidence that by 6 months, 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 infants develop the capacity to see actions as goal-directed.

Woodward (2013) suggested that infants come to perceive an action as goal-directed after they have intentionally performed that action with a particular goal in mind (e.g., for goal-directed reaching, at 6 months of age). At 12 months, infants also understand that individual actions can be related to each other based on their role in bringing about some overarching goal. 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 physical and psychological constraints of the action context), then the novel actions are perceived as intentional means to the familiar end (Woodward & Sommerville, 2000).

This is a plausible suggestion for how infants infer the goal of a novel action in cases where a single 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 simultaneous but distinct 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 principle drawing together three aspects of reality: actions, goal states, and situational constraints (Csibra, Bıró, Koós, & Gergely, 2003; Csibra & Gergely, 1998; Gergely & Csibra, 2003; see also Scott & Baillargeon, 2013). The principle of rational action relates these elements into a teleological schema by assuming that actions realize goal-states in the most efficient way possible. For example, in one classic experiment, 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, & Bíró, 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 that action is an efficient means to bring about any of the effects that have been observed. However, as Csibra and Gergely (2007) pointed out, there are cases where more than one effect is brought about in the most efficient manner. In case of this sort, their schema does not tell us how we infer which of the effects was intended. Instead, they admit that additional cognitive constraints are needed to determine the goal of the action.

Biro and 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 (Di Giorgio, Lunghi, Simion, & Vallortigara, 2017). A domain-specific learning mechanism can then detect statistical regularities about the surface-level features of the objects that typically exhibit the cues, for example, hands. Once agents are identified, infants can learn to infer their goals by keeping track of what effects those agents typically cause (and reasoning that effect typicality predicts goal likelihood). Critically, this mechanism does not address the question of how infants (or adults) might determine the goal of a novel action observed for the first time; it only addresses how infants can determine the familiar goals of agents.

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 (cf. Woodward, 2013). This account finds neuroscientific support in the possible role of mirror neurons in understanding others' minds (e.g., Gallese & Goldman, 1998). However, 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).

Extending Premack (1990), Baron-Cohen (1997) suggests that early in infancy humans may be equipped to recognize certain predefined goals such as freedom, companionship, arousal, arriving at a certain endpoint, affecting another agent, and reciprocating from a previous interaction. Nevertheless, this theory likewise does not have a clear method of dealing with novel actions that fall outside the schemas infants are prepared to deal with.

In sum, theories of intention inference have found it difficult to explain how observers infer the goal of a novel action with multiple salient effects. As a solution to this problem, we propose that domain-specific prior knowledge may be critical. This knowledge could take the form of innate constraints (e.g., Mikhail, 2011; Spelke, Bernier, & Skerry, 2013) or of prior distributions over the effects that are more likely to be intended. In the latter case, individuals 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).

There is already some evidence to suggest that domain-specific priors may help infants disambiguate the goals of novel actions with multiple salient effects. For example, Sommerville and 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 can interpret this action in at least two distinct ways: as directed toward the obtaining the toy or as directed toward obtaining the cloth (Sommerville & Woodward, 2005). The insight of Sommerville and Crane (2009) is that 10-month-olds could be encouraged to interpret the cloth-pulling behavior as being directed toward the toy if they were previously shown the experimenter reaching for and obtaining that toy in a nonambiguous context. These data make the 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 toward 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 informed about the goals of a particular agent to infer that agent's goal? It seems plausible that 12-month-olds already have a prior expectation about agents and out-of-reach objects and that the prior can now be applied to any agent. This prior knowledge allows 12-month-olds to solve the inference problem for novel actions with multiple salient effects in this narrow context (namely, reaching for objects). Our suggestion is that a process analogous to this may occur in the moral domain. On this view, priors are built up over the probability of an actor intending a good or bad effect. Alternatively, some priors may be "built in" or part of core cognition.

Positing a Solution

In sum, we suggest that the general problem of inferring the goal of a novel action with multiple effects can be solved with domain-specific prior knowledge. In the next section, we propose a specific solution to the problem of goal inference for a novel action with multiple salient effects. Just as infants can learn that object-obtaining is a more likely goal than cloth-pulling and can use this information to disambiguate the goals of agents, we propose that prior knowledge in the moral domain can similarly be used to disambiguate morally charged actions with multiple salient effects. In particular, we suggest that when there are two morally charged effects, one good and one bad, this prior knowledge favors the inference that the agent intended the good effect and not the bad effect (Mikhail, 2007, 2011). Put another way, a unique solution to the goal inference problem for novel actions in the moral domain can be achieved by positing a particular type of domain-specific knowledge: a good intention prior.[1]

Continuity of Infant, Preschool, and Adult Intention Inference Processes

Until now, we have been discussing how infants before the first year of life infer the intention of novel actions with multiple salient effects. Interpreting novel actions is arguably a more significant problem for infants than it is for children and adults (who see fewer novel act-types each day than infants), which partially explains why the literature on novel action inference is concerned with infants. However, when a novel action with multiple salient effects *is* seen by a child or an adult, the puzzle of how they infer the goal of that action still exists.

No mental machinery has yet been posited to explain how this puzzle might be solved in adult cognition differently than in infant cognition. In fact, many theories of intention inference stress the continuity between the infant and adult capacities. Some theories 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, 2005) 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. 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). Yet, the problem of how adults and children infer the goal of a novel action with multiple salient effects remains. As discussed below, this problem is particularly critical for the study of moral cognition in adults and children.

Goal Inference in Moral Cognition

Interpreting the goal of an action is critical to making a moral judgment about that action and arguably even more so for novel actions. For actions that have both morally good and morally bad effects, the agent's intention concerning those effects may be

[1] Although we will use the term "good intention prior" throughout this article, what we mean more precisely is "prior that an agent intends the good effect." We do not take a position here on whether intending the good effect amounts to having a "good intention" or whether other requirements are necessary for an intention to carry that distinction. In addition, although we posit the use of domain-specific prior knowledge to solve the problem of inferring the goal of a novel action with multiple salient effects, in this article we remain agnostic about where one domain starts and another begins and even what counts as a domain at all. Therefore, although we will refer to our hypothesis as being specific to the "moral domain," it is possible that the proper domain of the good intention prior is in fact broader. For instance, this prior may apply to the entire evaluative domain (containing the subdomains of pragmatics, aesthetics, and so forth). In contrast, it is also possible that the good intention prior exists in a narrower domain than the moral domain, for instance, the domain of deontic judgments of harm (a subdomain of the moral domain).

necessary to determine the moral permissibility of the action. Many theories of moral cognition highlight the role that intention plays in making moral evaluations for adults (Cushman, 2013; Greene, 2013; Malle, Guglielmo, & Monroe, 2014; Mikhail, 2011; Young et al., 2007; for a review, see Doris & Moral Psychology Research Group, 2010), as well as children (Baird & Astington, 2004; Cushman, Sheketoff, Wharton, & Carey, 2013; Killen et al., 2011; for a review see Killen & Smetana, 2015), and 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).

Trolley Problems as a Test Case

The literature on moral psychology has made good use of a certain kind of moral dilemma, often termed "the trolley problem," to test how certain features of moral perception—such as intention, outcome, and causal sequence—impact the moral permissibility of actions (Cushman, Young, & Hauser, 2006; Greene, Cushman, Stewart, Lowenberg, Nystrom, & Cohen, 2009; Mikhail, 2002; Pellizzoni, Siegal, & Surian, 2010; Schwitzgebel & Cushman, 2012; Waldmann & Dieterich, 2007; for a review, see Waldmann, Nagel, & Wiegmann, 2012). Trolley problems also raise the puzzle of how we discern the goal of a novel action that has more than one salient moral effect (Mikhail, 2007, 2011). 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 is a novel action that leads to two salient moral effects: some people are saved while others are killed. How do subjects who are only given information about the causal sequence of events that occur infer the intention of the agent? Was the intention to save lives, or cause deaths, or both? (See Figure 1). Do subjects choose randomly between possible intentions for the agent? Or are they consistent in how they determine the agent's intention? This question is highlighted in the trolley problem scenario but arises for any case of action that involves multiple morally charged effects.

Because most theories of goal inference fall short of being able to explain how we infer the goal of a novel action with multiple salient effects, moral psychologists cannot simply "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.

Theories of Moral Cognition Fail to Address the Goal Inference Problem

Although some researchers who emphasize the role of intention in moral judgment have simply avoided the question of how intention is inferred (often explicitly telling subjects what an agent intends; e.g., Baird & Astington, 2004; Cushman et al., 2013; Young & Saxe, 2011), others have attempted to describe how action representations are built from impoverished stimuli (including stimuli lacking explicit intention information) so that moral judgment can proceed. While not always acknowledging the difficult nature of the goal inference problem, this second group of investigators has either explicitly or tacitly assumed something like a good intention prior.

Mikhail (2000, 2007, 2009, 2011) was the first cognitive scientist to highlight the fact that trolley problems can be used as a tool to investigate the goal inference problem. He pointed out that computing the intention of the agent on the basis of an impover-

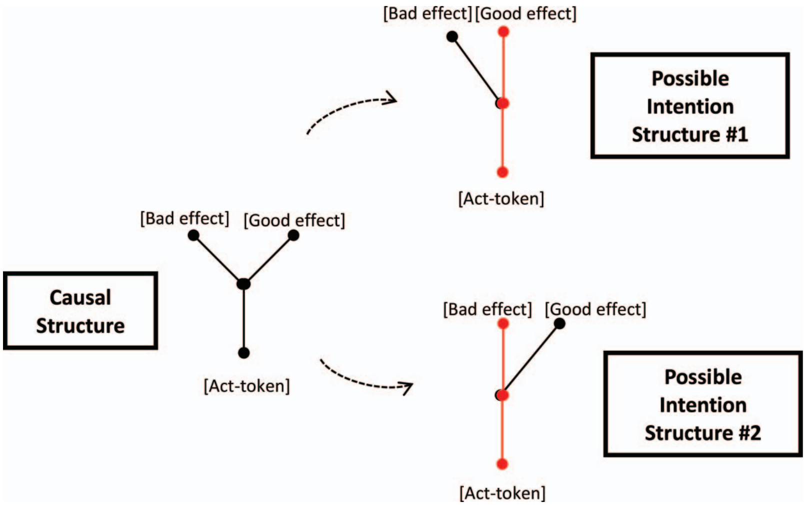
[Bad effect] [Good effect]
Possible Intention Structure #1
[Act-token]
[Bad effect] [Good effect]
Causal Structure
[Act-token]
[Bad effect] [Good effect]
Possible Intention Structure #2
[Act-token]

Figure 1. When one action has two effects (e.g., one good effect and one bad effect), a single causal structure is compatible with at least two different intention structures. In this figure, the red (dark grey/vertical) line indicates the agent's action plan—the sequence of actions she intends to bring about her goal. 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, however, is also a viable transformation of the causal structure (Possible Intention Structure #2). See the online article for the color version of this figure.

ished stimulus that lacks any goal information requires making assumptions about which of the morally good or bad effects the agent intends. Mikhail hypothesized that subjects do this by way of the good intention prior, or what he figuratively called "a presumption of innocence" (e.g., Mikhail, 2009, p. 90). Until now, however, this proposal has lacked adequate empirical support.

Although no other theory has made this proposal explicit, several seem to tacitly assume a good intention prior. For example, Greene (2013) suggests that individuals 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 action plans, Greene does not provide an account of how subjects infer which effects count as side effects, means, and goals. However, each of the action plans he uses to illustrate how particular dilemmas are represented assumes that saving lives is the goal of the agent (e.g., see Figures 9.7-9.10 in Greene, 2013). In the background of Greene's theory, therefore, is the critical, untested assumption that agents are presumed to intend the good effects and not the bad effects.

Two recent theories (Crockett, 2013; Cushman, 2013) suggest that mechanisms of moral judgment can be described using a dual-process approach that is instantiated by "model-free" and "model-based" reinforcement learning systems. These theories suggest that a model-based system calculates the expected outcome of a moral action and places value on the outcome. The model-free system places value on moral actions (such as pushing) as well as on subgoals. This role of the model-free system allows for intentions to be represented as multistep plans in a hierarchical goal/subgoal structure. A combination of the outputs of the two systems produces a moral judgment. However, there is no clear method for the system to independently determine what the goal of the agent is in the first place. For this process to get off the ground, a background assumption must be made that the agent acts "out of concern for others rather than malice" (Cushman, 2013, p. 283). Given this assumption, goals and subgoals can be assigned to the agent and moral cognition can proceed. Again, that assumption remains untested.

Experiment 1

Our experiments use trolley problem scenarios because of the clarity with which they present subjects with a case that has two effects, one good and one bad, which are precipitated by a novel action. We hypothesize that in the traditional trolley case, when no intention information is explicitly given (the Uninformative Condition), subjects will use the good intention prior to determine which of the effects were intended. In particular, they will judge the good effect to be intended and the bad effect not intended.

We will compare the Uninformative Condition to two separate "informative" cases. In the first informative case, we will explicitly tell subjects that the agent intended the good effect and not the bad effect (Informative Condition Good); we expect this to not make a difference to judgments of intention or moral permissibility. That is, we expect there to be no difference in how subjects judge the cases whether we tell them nothing about the agent's intention or we explicitly tell them that the agent intended the good effect and not the bad effect. We suggest that this provides evidence that the "additional" information (the information explicitly given) about the agent's intention is not additional information at all—rather, subjects use that information as a prior when no information is explicitly given. To test this hypothesis that the Uninformative Condition and Informative Condition Good are not different (on measures of intention ascription and moral judgment), we will use a Bayesian analysis to weigh evidence for the null hypothesis.

In the second informative condition, we explicitly tell subjects that the agent intended the bad effect (Informative Condition Bad). We expect subjects' judgments in this case to be different from those of the Uninformative Condition (where no information was explicitly given) because this additional information is actually additional: It provides information about the agent's intention that the subject would not have otherwise assumed.

In sum, the central analysis will be to determine whether subjects' judgments in the Uninformative Condition look more like judgments in the Informative Condition Good or the Informative Condition Bad. We hypothesize that intention and permissibility judgments of the Uninformative Condition will look like judgments in the Informative Condition Good. However, it is also possible that, in the Uninformative Condition (when no intention information is explicitly given), some subjects will impute good intentions to the agent and some subjects will impute bad intentions. A third possibility is that most of the subjects will impute bad intentions. We differentiate between these possibilities in our analysis (described in more detail below).

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 the Appendix). The causal structure of the agent's action was presumed to be unambiguous.[2] However, (at least) 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). (See Figure 2).

Subjects were randomly assigned to receive the story in one of three conditions. Subjects in the Uninformative Condition received no explicit information about the agent's intention. Subjects in the Informative Condition Good received information that the agent intended the good effect of his action (saving the five people on the main track). Subjects in the Informative Condition Bad 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 answering the following question: "Why do you think Hank threw the switch?" Subjects chose between two answer choices presented in randomized order: "To save the five men on

[2] Strictly speaking, the causal structure is not entirely unambiguous, insofar as background assumptions are also necessary to compute causal structures. However, we set this issue aside for the purposes of this investigation.

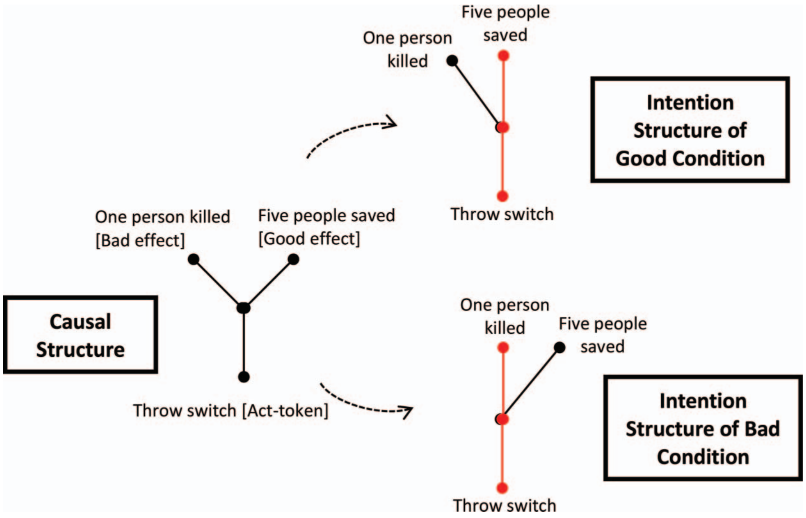
Five people saved
One person killed
Intention Structure of Good Condition
Throw switch
One person killed [Bad effect]
Five people saved [Good effect]
Causal Structure
Throw switch [Act-token]
One person killed
Five people saved
Intention Structure of Bad Condition
Throw switch

Figure 2. The relationship between the causal structure and the intention structures of the story in Experiment 1. The causal structure is compatible with (at least) two intention structures. In the Informative Condition Bad and Informative Condition Good, information is provided which allows subjects to choose an intention structure. In the Uninformative Condition, no intention information is explicitly given. See the online article for the color version of this figure.

the main track" or "To kill the one man on the side track."[3] Study materials are available at https://github.com/sydneylevine/good-intention-prior.

Subjects. The study procedure was approved by the Committee on the Use of Human Subjects in Research at Harvard University. Subjects were recruited from Amazon's Mechanical Turk platform and were paid for their participation. We stopped data collection when 180 adult subjects completed the study to achieve sample sizes comparable to previous studies that have used similar methodology (cf. Greene et al., 2009), that is, sizes of about 60 subjects per condition. Because of an error in the randomizer, 63 subjects received the Uninformative Condition, 58 received the Informative Condition Good, and 59 received the Informative Condition Bad. Eleven subjects were excluded from analysis for failing an attention check, leaving 60 subjects in the Uninformative Condition, 54 in Informative Condition Good, and 55 in Informative Condition Bad.

Statistical Analyses. Our hypothesis is that in the Uninformative Condition, where the agent's intention is not specified, subjects will assume that the agent's intention is to bring about the good effect. The consequence of this is that intention inferences will be the same in the Uninformative Condition as in the condition where good intentions are stipulated (Informative Condition Good). Moreover, because we assume that inferred intention is an important determinant of moral judgments, we further hypothesized that moral judgments in the Uninformative Condition will be the same as judgments in the Informative Condition Good. Conventional statistics are not suited to the assessment of this hypothesis, because it is a null hypothesis. Conversely, we also hypothesized that the Uninformative Condition will be significantly different than the condition where bad intentions are stipulated (Informative Condition Bad) on measures of both intention inference and moral judgment. This hypothesis is more suited to conventional statistics, because it makes a prediction about rejecting the null hypothesis.

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. It is therefore possible to compute the relative likelihood of the competing hypotheses given the data (the Bayes factor). We computed both Bayes factors (using a code written by Randy Gallistel and available online here: https://github.com/sydneylevine/good-intention-prior/blob/master/BinoBF2_commented.m) and conventional *p* values in deference to current common statistical practices, though our main conclusions are primarily drawn from the Bayesian analysis.

In computing Bayes factors, we considered two alternatives to our null hypothesis: The first is that the probability of a given judgment in the Informative Condition Good provides no information about judgments in the Uninformative Condition. On this alternative, the probability of a judgment in the Uninformative 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 two-tailed *t* test for difference in the means. A more refined alternative is that the judgments and ratings in the Uninformative Condition will be more negative than in the Informative Condition Good, 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 Uninformative Condition may with equal probability assume any value on the negative side

[3] A slightly different version of Experiment 1 is described in the online supplementary materials. The results of that experiment, also reported in the online supplementary materials, replicate the results reported here.

of the value in the Informative Condition Good. This is the simplest formulation of what the implicit alternative is in a one-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 = trivial support; 2 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.

Proving the null requires more data because the null is a point hypothesis and point hypotheses are much stronger hypotheses than interval hypotheses, for the simple reason that any interval hypothesis subsumes an uncountable infinity of point hypotheses (all the points within the interval). It is often argued that null hypotheses are so strong that they can never actually be true, because for trivial, uninteresting reasons, there will always be some difference no matter how minute (see Morey & Rouder, 2011 and citations therein). One never actually proves any hypothesis with a statistical hypothesis evaluation. In a p test, one computes how improbable the observed outcome would be under the null hypothesis. This, of course, says nothing about how improbable some specified alternative to the null is. In an null hypothesis statistics test, an alternative is never specified; a fortiori, its probability is never computed.

A Bayesian analysis computes the relative likelihoods of two hypotheses given the data. One of these is usually the null. Because hypotheses are unlike outcomes in that they are neither mutually exclusive nor exhaustive, it is possible for a null hypothesis to be more likely than a set of hypotheses of which the null is a member (the uncountably infinite set of hypotheses falling within an interval that includes the null). However, the relative likelihood of the null can only be higher than the set that includes it if the mode of the likelihood function is very close to the null and the likelihood function is much narrower than the interval specified by the alternative to the null. A narrow likelihood function can only be obtained with a lot of data. By contrast, any likelihood function whose mode lies well away from the null will yield a high Bayes Factor in favor of the alternative (that is, in favor of the interval, rather than the null point within that interval; see Morey & Rouder, 2011, for further discussion).[4]

Results

First, we will consider subjects' intentionality judgments, the more critical measure for providing support for the good intention prior. In the Uninformative Condition, 98.3% of subjects (59 out of 60) judged that the agent intended the good effect of his action. Likewise, in the Informative Condition Good, 98.1% of subjects (53 out of 54) judged that the agent intended the good effect of his action. In the Informative Condition Bad, 32.7% of subjects (18 out of 55) judged that the agent intended the good effect of his action.

Critically, there was no significant difference in intention judgments between the Uninformative Condition and Informative Condition Good, Upton's $\chi^2(1, N = 114) = .0056$, $\varphi = .0071$, $p = .95$, two-tailed. In addition, the Bayes factors favored the null (14.76 two-tailed, 28.6 one-tailed). In contrast, there was a significant difference in intention judgments between the Uninformative Condition and Informative Condition Bad, Upton's $\chi^2(1, N = 115) = 55.34$, $\varphi = .69$, $p < .0001$, two-tailed, with Bayes factors decisively in favor of the alternatives to the null, whether one- or two-tailed (see Figure 3).

Next, we will consider subjects' deontic judgments. In the Uninformative Condition, 71.7% of subjects (43 out of 60) judged the case permissible. In the Informative Condition Good, 77.8% of subjects (42 out of 54) judged the case permissible. In the Informative Condition Bad, 49.1% of subjects (27 out of 55) judged the case permissible.

There was no significant difference in permissibility judgments between the Uninformative Condition and Informative Condition Good, Upton's $\chi^2(1, N = 114) = .55$, $\varphi = .069$, $p = .46$, two-tailed. The Bayes factors favored the null (3.79 two-tailed; 3.68 one-tailed). In contrast, there was a significant difference in intention judgments between the Uninformative Condition and the Informative Condition Bad, Upton's $\chi^2(1, N = 115) = 6.09$, $\varphi = 0.23$, $p = .013$, two-tailed, with Bayes Factors decisively in favor of the alternative to the null whether it was one- or two-tailed (see Figure 4).

Discussion

The main finding of Experiment 1 is that subjects in the Uninformative Condition infer that the agent intended the good effect of his action when no intention information was explicitly given. In fact, subjects' judgments in the Uninformative Condition are the same as subjects' judgments in the Informative Condition Good, in which they are explicitly told that the agent intended the good effect (Bayes factors provided support for the null hypothesis). These findings suggest that our hypothesis is correct: subjects in the Uninformative Condition use the good intention prior to determine the intention of the agent.

Although we did not explicitly give any intention information to subjects in the Uninformative Condition, it is possible that there is a salient clue to intention in the way the traditional trolley problem is arranged, with five people on the main track who are threatened by the train and one person on the side-track. Assume that the following conditions hold: first, that subjects come to the scenario with no priors about whether the agent intends to maximize people saved (generally intends good effects) or people killed (generally intends bad effects). Second, also assume that if the agent does have people in the world that he would be interested in killing or saving, that it would be incredibly unlikely for them to be standing on the tracks in just the way that would allow the agent to kill or save them. Put another way, if the agent is, in fact, interested in killing someone, the odds are small that he would appear on the side-track and not on the main track or somewhere else entirely. Given these assumptions, then the fact that the agent flips the switch provides evidence that the agent intends to maximize lives saved as a general policy. After all, if the agent intended to

[4] We are grateful to Randy Gallistel for his help with this section.

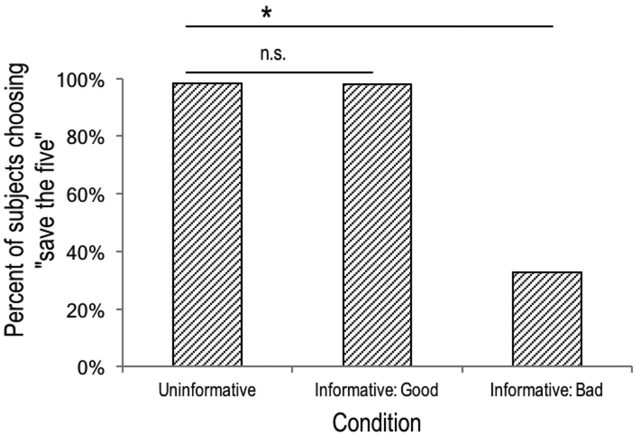

Figure 3. Adult subjects' judgments of the intention of the agent in Experiment 1. Subjects responded to the question, "Why do you think Hank/Joe/Mark threw the switch?" and selected from two options: "To save the five men on the main track" and "To kill the one person on the side track". * $p < .05$.

maximize death, he would have stood there, done nothing and happily watched the train run over the five people on the main track. (On this view, the flipping of the switch should not be interpreted as an agent intending to save those five particular people or kill that one particular person, because the odds of having attachments to those people is so small). The subject then uses the information that the agent generally intends good effects to infer that the agent intended to save the particular five people on the track, rather than to kill the particular one person. Experiment 2 was designed to address this concern.

Experiment 2

Methods

Experimental design. Subjects read a story in which a train is about to kill one person who is standing in its path. In response, an agent throws a switch, thereby preventing the train from killing the one person and with the same action causing the train to turn down a side-track and kill five people. (For text of the stimuli, see the Appendix). As in Experiment 1, at least two intention structures are compatible with the causal structure: it is possible that the agent's intention was to save the one person or that the agent intended to kill the five people. If subjects in Experiment 1 made their intention inference based on the fact that more people would be saved by flipping the switch then by doing nothing (and that therefore the subject intends to maximize lives saved), then subjects should infer in this case that the agent intended to kill the five because in this case, the agent's action maximizes harm.

Subjects were randomly assigned to receive the story in one of three conditions: uninformative, Informative Condition Good, and Informative Condition Bad. These were identical to the conditions of Experiment 1, except that one person was on the main track and five people were on the side track. Subjects were asked the same two test questions as in Experiment 1, the moral permissibility question ("Is it morally permissible for Hank/Joe/Mark to throw the switch?") and the intention question ("Why do you think Hank/Joe/Mark threw the switch?"). The two options for the intention question in this experiment were "To save the one man on the main track" and "To kill the five men on the side track." Study materials are available at https://github.com/sydneylevine/good-intention-prior.

Subjects. The study procedure was approved by the Committee on the Use of Human Subjects in Research at Harvard University. Subjects were recruited from Amazon's Mechanical Turk platform and were paid for their participation. We stopped data collection when 181 adult subjects completed the study, to approximate the sample sizes of Experiment 1. Sixty-one subjects received the Uninformative Condition, 60 received the Informative Condition Good, and 60 received the Informative Condition Bad. Sixteen subjects were excluded from analysis for failing an attention check, leaving 57 subjects in the Uninformative Condition, 51 in Informative Condition Good, and 57 in Informative Condition Bad.

Results

In the Uninformative Condition, 94.7% of subjects (54 out of 57) judged that the agent intended the good effect of his action (saving the one man on the main track). Likewise, in the Informative Condition Good, 98.0% of subjects (50 out of 51) judged that the agent intended the good effect of his action. In the Informative Condition Bad, 17.5% of subjects (10 out of 57) judged that the agent intended the good effect of his action.

Critically, there was no significant difference in intention judgments between the Uninformative Condition and Informative Condition Good, Upton's $\chi^2(1, N = 108) = .82$, $\varphi = .087$, $p = .37$, two-tailed. In addition, the Bayes factors favored the null (7.75 two-tailed, 9.27 one-tailed). In contrast, there was a significant difference in intention judgments between the Uninformative Condition and Informative Condition Bad, Upton's $\chi^2(1, N = 114) = 68.37$, $\varphi = .77$, $p < .0001$, two-tailed, with Bayes factors decisively in favor of the alternatives to the null, whether one- or two-tailed (see Figure 5).

With respect to permissibility judgments, in the Uninformative Condition, 29.8% of subjects (17 out of 57) judged the case permissible. In the Informative Condition Good, 35.3% of subjects (18 out of 51) judged the case permissible. In the Informative

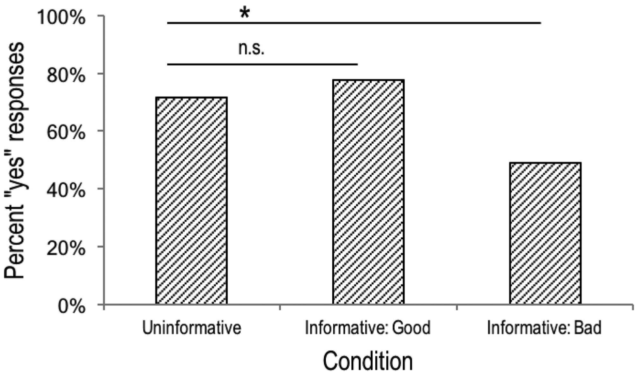

Figure 4. Adult subjects' deontic judgments in Experiment 1. Subjects answered the question, "Is it morally permissible for Hank/Joe/Mark to throw the switch"? * $p < .05$.

Condition Bad, 17.5% of subjects (10 out of 57) judged the case permissible.

There was no significant difference in permissibility judgments between the Uninformative Condition and Informative Condition Good, Upton's $\chi^2(1, N = 108) = .36$, $\varphi = .058$, $p = .55$, two-tailed. The Bayes factors favored the null (3.76 two-tailed; 1.70 one-tailed). There was a marginally significant difference in permissibility judgments between the Uninformative Condition and the Informative Condition Bad, Upton's $\chi^2(1, N = 114) = 2.36$, $\varphi = 0.14$, $p = .12$, two-tailed. The Bayes factor was not decisive in this case (1.59 two-tailed, 1.36 one-tailed, see Figure 6).

Discussion

As in Experiment 1, subjects in Experiment 2 inferred that the agent in the Uninformative Condition intended the good effects of his action, despite the fact that it did not maximize lives saved. If subjects entered this case with no priors about whether the agent generally intends good effects or bad effects, then observing the agent flip the switch and kill the five men should provide evidence that he generally intends bad effects. Instead, subjects infer that the agent intended to save the one man on the main track. Because this is the same judgment subjects give when the agent's intention to bring about the good effect is explicitly stated, it suggests that subjects in the Uninformative Condition approach the problem with a good intention prior.

The main difference in the results of Experiment 1 and Experiment 2 was subjects' permissibility judgments. In the uninformative and informative good conditions of Experiment 1, over 70% of subjects morally approved of the agent's action of flipping the switch, whereas only about 30% of subjects approved of the agent's action in those conditions in Experiment 2. Put simply, in Experiment 2, when the agent redirects the train away from the one person toward the five, even though subjects think that the agent intended the good effects of his action, they think the action was morally impermissible. This underscores the point that while intention makes a difference for moral judgment, it is not the sole metric upon which moral actions are evaluated. The agent's action in the original version of the side-track trolley problem (the Uninformative Condition in Experiment 1) is judged permissible not only because the agent intends the good effects but also because the good effects outweigh the bad effects (Mikhail, 2011), a requirement that does not hold in Experiment 2.[5]

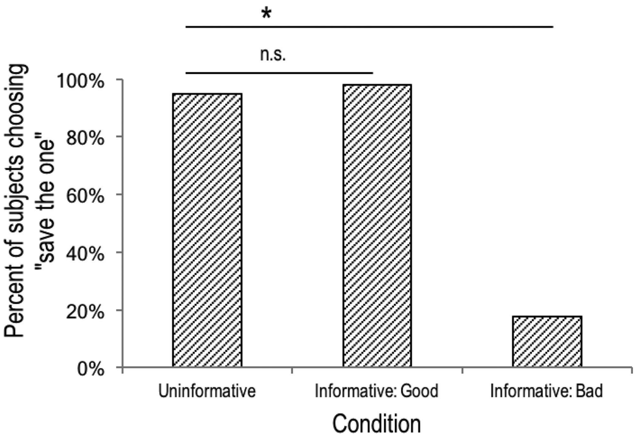

Figure 5. Adult subjects' judgments of the intention of the agent in Experiment 2. Subjects responded to the question, "Why do you think Hank/Joe/Mark threw the switch?" and selected from two options: "To save the one man on the main track" and "To kill the five men on the side track". * $p < .05$.

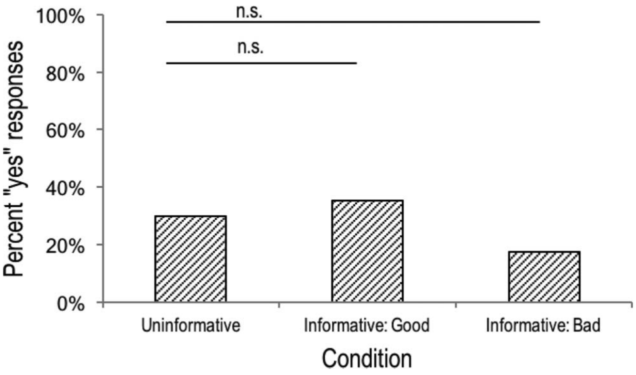

Figure 6. Adult subjects' deontic judgments in Experiment 2. Subjects answered the question, "Is it morally permissible for Hank/Joe/Mark to throw the switch"?

Experiment 3

Does the good intention prior take decades of social learning to emerge in adults? Or is it present already by the preschool years? Experiment 3 was designed to test whether preschoolers, like adults, use the good intention prior when confronted with a case of a novel action with two salient effects.

Methods

Experimental design. Subjects were tested individually in quiet locations in their preschools or in the lab. Following Leslie, Mallon, and DiCorcia (2006), subjects were first trained on use of a Likert scale, the "pink 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 the Appendix 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 impermissi-

[5] In Mikhail's (2011) view, there are at least two other factors present in the trolley problem that make that action of foreseen harm permissible: that there is no morally preferable alternative and that the act itself is not wrong.

ble." Although, our data suggest that children respond to the question "should she have done that" in a similar way that adults respond to straightforward questions of deontic judgment, it remains an open question whether "should" captures a 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 were randomly assigned to receive the story in one of three conditions. Subjects in the Uninformative Condition received no explicit information about the agent's intention. Subjects in the Informative Condition Good received information that the agent intended the good effects of her action (saving the five children's cookies). Subjects in the Informative Condition Bad 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 the Appendix for full text of each story and control questions). 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?"[6] 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?" Study materials are available at https://github.com/sydneylevine/good-intention-prior.

Subjects. The study procedure was approved by the Rutgers University Institutional Review Board for the Protection of Human Subjects. Fifty children between the ages of 37 months and 72 months received the Uninformative Condition (M = 55.6 months; SD = 9.0 months), 29 of which were girls. Forty-seven children between the ages of 40 months and 72 months received the Informative Condition Good (M = 53.8 months; SD = 8.2 months), 23 of which were girls. Thirty-six children between the ages of 42 months and 68 months received the Informative Condition Bad (M = 56.3; SD = 8.1), 17 of which were girls. 37 additional children were excluded from the study, 32 for failing scale training, two for failing to cooperate, one for failing control questions, one for parent interference, and one for experimenter error.

As is accepted (and even encouraged) in the Bayesian tradition, sample sizes were not preset. Thirty-six subjects were collected in each condition (approximately the size of samples used for similar previous studies, e.g., Saunders, 2014) and then the Bayes factors were calculated for the contrasts of interest (Uninformative Condition compared to Informative Condition Good and Uninformative Condition compared to Informative Condition Bad). The Bayes factor was decisive for the latter contrast, so no further data was collected in the Informative Condition Bad. Bayes factors were uninformative in the former contrast, so data collection continued. Despite the fact that "optional stopping" is a major concern for null-hypothesis statistics testing and allows for extremely problematic "p-hacking," there is no such concern when Bayesian analysis is used. (For extensive treatment of this issue, see Edwards, Lindman, & Savage, 1963; Rouder, 2014; Wagenmakers, 2007; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). Our main conclusions were based on the results of the Bayesian analysis, though we also compute conventional p values in deference to current common statistical practices and for comparison with other work.

Results

First, we will consider subject's intentionality judgments. In the Uninformative Condition, 66% of subjects (33 out of 50) judged that the agent did not intend the bad effect of her action. In the Informative Condition Good, 60% of subjects (28 out of 47) judged that the agent did not intend the bad effect of her action. In the Informative Condition Bad, 14% of subjects (five out of 36) judged that the agent did not intend the bad effect of her action.

Critically, there was no significant difference in intention judgments between the Uninformative Condition and Informative Condition Good, Upton's $\chi^2(1, N = 97) = .42$, $\varphi = .07$, $p > .25$, two-tailed, and the BFs show that the results favor the null hypothesis (7.20 one-tailed, 3.37 two-tailed). In addition, there was a significant difference between the Uninformative Condition and Informative Condition Bad, 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 one- or two-tailed (see Figure 7).

Next, we will consider subjects' deontic judgments. In the Uninformative 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 Informative Condition Good, 72% of subjects (34 out of 47) judged the case permissible. In the Informative Condition Bad, 22% of subjects (eight out of 36) judged the case permissible.

There was no significant difference between subjects' responses to the Uninformative Condition and Informative Condition Good, Upton's $\chi^2(1, N = 97) = 1.62$, $\varphi = .13$, $p = .202$, two-tailed. In contrast, there was a significant difference between the Uninformative Condition and Informative Condition Bad, Upton's $\chi^2(1, N = 86) = 11.96$, $\varphi = .37$, $p < .001$, two-tailed. The one-tailed BF for the good–uninformative comparison was 1.47 in favor of the null; the two-tailed BF was 1.90 in favor of the null. The BFs for the good–bad and uninformative–bad comparisons were all decisive (>100) in favor of the alternative hypothesis (see Figure 8).

Next, we will consider subjects' Likert ratings of the action of the agent. Likert scale ratings were ranged from −2 (*really bad*) to 2 (*really good*). Subjects in the Uninformative Condition and the Informative Condition Good both rated the agent's action as slightly above the midpoint of the scale (uninformative: $M = .24$, $SD = 1.36$; good: $M = .23$, $SD = 1.31$). Subjects in the Informative Condition Bad 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

[6] We follow Leslie et al. (2006) in using the "on purpose" question to measure intention inference in our preschool subjects.

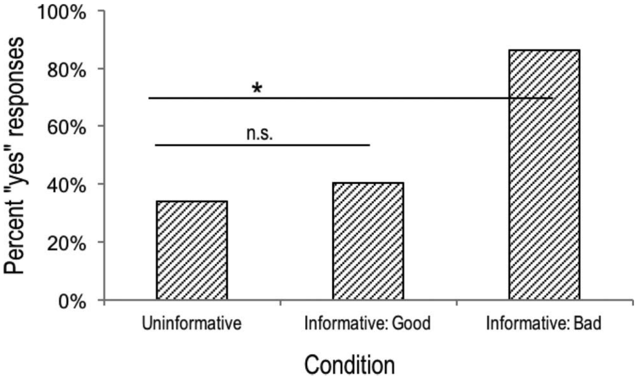

Figure 7. Preschool subjects' judgments of the intention of the agent in Experiment 3. Subjects answered the question, "Did she make the one kid sad on purpose"? * $p < .05$.

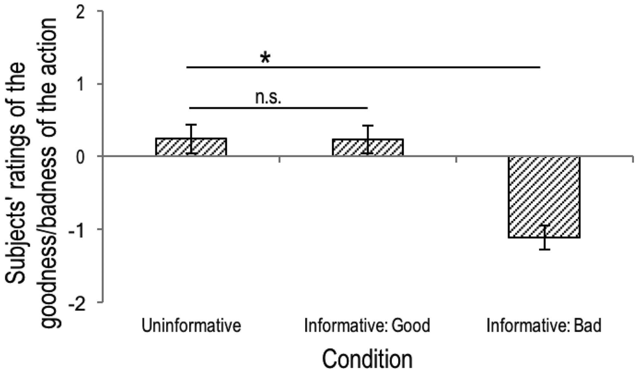

Figure 9. Subjects' ratings of the action of the protagonist in Experiment 3. Subjects responded to the question, "Was what Sally did good, bad, or just OK?" Likert scale ratings ranged from −2 (*really bad*) to 2 (*really good*). Error bars show standard error of the mean. * $p < .05$.

revealed that there was no significant difference between the Uninformative Condition and Informative Condition Good (independent-sample t test, two-tailed, $t(95) = .022$, $r = .0022$, $p > .250$). Furthermore, the two-tailed BF for the good–uninformative comparison was 5.65 in favor of the null. In contrast, there was a significant difference between the Uninformative Condition and the Informative Condition Bad (independent-sample t test, two-tailed, $t(84) = 5.24$, $r = .50$, $p < .001$). The BFs for the good–bad and uninformative–bad comparisons were all decisive (>100; see Figure 9).

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.

We will now consider a possible objection to our interpretation of the results of Experiment 3. Although we did not provide any explicit intention information in the Uninformative Condition, it seems conceivable that incidental features of the wording of the story may have impacted subjects' inferences. For example, subjects hear about the prospect of the good effect (that the five kids will be blocked by the gate) in the moment that the protagonist acts, and only afterward hear about the impending bad effect (that the one child's cookie is about to be eaten). Is it possible that this sequence is enough to bias the subject toward inferring that the agent intended one effect and not the other? There are a few reasons that this explanation seems unlikely.

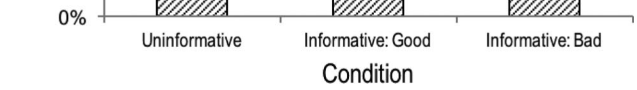

Figure 8. Preschool subjects' moral judgment of the action of the agent in Experiment 3. Subjects answered the question, "Should she have done that"? * $p < .05$.

First, the good and bad effects in which we are interested (and the experimental effects that our dependent variable queries) relate to the ultimate outcomes of the scenario, that one kid is happy and five kids are sad. These results are causally downstream from the events that happen when Sally puts up the gate, namely, that the squirrel is (a) blocked and (b) redirected. It is true that subjects hear about the squirrel being blocked before they hear about the redirection. However, they hear about the actual occurrence of the effects in the reverse order—first hearing that the one kid is sad and only then that the others are not sad. Second, when subjects recount the events of the story during the second telling, they are asked to explicitly report what the squirrel does when the protagonist puts up the gate. The correct answer is that the squirrel ate the one kid's cookie. In this way, subjects are asked to actively report the connection between the protagonist's action and the bad effect, which in some ways may be thought to skew their attention toward the bad effect. Following this, subjects are then asked to first report how the one kid feels (sad) and only then how the five kids feel (not sad). In sum, in three of the four ways in which the outcomes are described in the experiment (prospect of the effects, occurrence of the effects, recounting the result of the protagonist's action, recounting the occurrence of the effects), the scenarios are balanced toward emphasizing the bad effect. It therefore seems unlikely that subjects were biased

Table 1
Correlation Matrix for Should, Likert Rating, and Intention Questions

Variable	Should	Action Rating	Intention
Should	—	.52**	−.22*
Action rating	—	—	−.32**

Note. Pearson correlations are reported. $n = 133$.
* $p = .011$, two-tailed. ** $p < .001$, two-tailed.

toward the good effect merely by the wording of the stories themselves.

General Discussion

Until now, empirical paradigms of intention inference generally have not been able to explain how individuals can infer the goal of a novel action when multiple salient effects are observed. We suggest that domain-specific prior knowledge can help solve the problem. If this is the case, then it is an empirical question what goals are favored by the prior knowledge. Furthermore, 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. Many theories of moral cognition, even those that highlight the important role that intention plays in moral judgment, have ignored this issue (e.g., Cushman et al., 2013; Young & Saxe, 2011) or have made tacit assumptions about subjects' intention inferences that have not been empirically validated (Crockett, 2013; Cushman, 2013; Greene, 2013). Furthermore, Mikhail's (2007, 2011) previous work on intention inference, the most significant and explicit exception to this pattern of neglect, does not supply systematic experimental evidence to support the theory it defends.

We propose an empirically testable solution to the problem of inferring intention for novel action in the moral domain: a good intention prior. In particular, we contend that when a novel action is observed that results in morally good and bad effects, then domain-specific prior knowledge favors the good effect as the actor's goal. If this hypothesis is correct, then when no intention information is available (e.g., in the Uninformative Condition), subjects should treat the morally good effect as the goal of the agent's action—just as they do in the Informative Condition Good. By contrast, if the hypothesis is incorrect, then subjects may (a) be more likely to treat the morally bad effect as the goal (as compared to the Informative Condition Good) or (b) choose equally between the two options.

The principle finding of our studies is that when no intention information is explicitly stated for a novel action (Uninformative Condition), adult and preschool subjects judge the case in the same way as cases in which they are explicitly told that the agent intends the good effect of the action (Informative Condition Good). This suggests that in our Uninformative 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 (Uninformative Condition and Informative Condition Good).

Our cases were designed to approximate the state of having no prior information about an agent's intention (besides what was explicitly stated in the stimulus). Learning additional information about a particular agent, his relationship to the other agents in the scenario, or his history and motives, of course, could all impact this prior knowledge or probability, ultimately leading an observer to attribute bad or ambiguous intentions to that agent. Moreover, it is worth noting that the subjects we tested generally came from relatively stable and nonviolent home environments (preschoolers were mostly residents of Middlesex County, NJ). Being exposed to more intentional harm on a regular basis could have the potential to influence individuals' prior assumptions about the intentions of agents in general.

Why the "Trolley" Task?

As we noted earlier, the "trolley" task is a useful empirical paradigm for our purposes because it sets up a single novel action that has both good and bad effects. Recently, some authors have argued that such tasks are of limited value because they lack external validity (Bauman, McGraw, Bartels, & Warren, 2014; Kahane, Everett, Earp, Farias, & Savulescu, 2015). Part of their concern is that the events depicted are "unrealistic" and subjects say they have never encountered this situation before. Certainly, the confluence of an out-of-control trolley, groups of people tied to railroad lines, certain death, and some of the other surface features of these tasks do seem unlikely. We simply make two points here. First, the novelty and unlikeliness of the events actually works in our favor insofar as we are interested in how subjects infer intention for novel actions. Second, even if one grants for the sake of argument that the general structure of such moral dilemmas is in some way atypical or occurs infrequently in everyday life, our larger aim is to differentiate between competing theories of underlying processes. As in other domains of cognitive science, pursuing this theoretical aim may require more than presenting subjects with everyday occurrences and may involve cases that are not directly informed by doctrine, politics, religion, or even conscious reasoning. (For further discussion of both points, see Mikhail, 2005).

A Good Intention Prior in Children? The Case For and Against

The case for. Margoni and Surian (2017) compared the age at which children rely on intention (as opposed to outcome) to make goodness and badness judgments for helping and harming scenarios, respectively. The authors found that by 4 years old, children judge cases of failed attempts at helping to be good, whereas it takes them until almost age 7 to judge failed attempts at harming to be bad. This suggests that, under some circumstances, children may be able to recognize the moral relevance of good intentions earlier than they recognize the moral relevance of bad intentions. This could explain why the preschool subjects in our studies were more likely to see the good effect as intended when the intention of the actor was ambiguous (as it was in the Uninformative Condition) and to use that information in their moral judgments.

The study by Margoni and Surian (2017) hints at the presence of a good intention prior in nontrolley setting. However, it remains an open question of how widely this prior is generalized. Put another way, what needs to be addressed is the proper domain of the good intention prior. Is there evidence for this prior across all moral subdomains (such as loyalty, sanctity, and authority; Graham, Haidt, & Nosek, 2009), or is it restricted to the harm domain or even to a subset of the harm domain? Alternatively, is it possible that the proper domain of this prior is a broader domain than the moral domain? Can it be found in other normative or evaluative domains (such as aesthetics, epistemology, economics, and so forth)? For example, 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). While this effect was first observed in morally-charged cases, we now know that the effect is not restricted to the moral domain but instead spans other evaluative domains (Knobe & Mendlow, 2004; MacHery, 2008; Rakoczy et al., 2015; Uttich & Lombrozo, 2010). Similarly, the good intention prior might also span other evaluative domains as well.

The case against. Two sets of recent findings may seem to contradict our proposal for a good intention prior. First, it has been suggested that there may be a "negativity bias" in agency attribution: When it is unclear if an object is an agent or a nonagent, 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). Together with our data, these findings suggest that there are separate cognitive processes at work for the attribution of intention, depending on whether the presence of an agent is certain or uncertain. In the latter 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 former case (as in our studies), when the presence of an agent is obvious 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. As mentioned above, 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, however, which may resolve the seeming contradiction. First, in side-effect effect cases, the protagonist in the story disavows one of the effects, that is, he explicitly declares that he does not care about it. Although 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 shown that such a statement indicates a certain kind of added intention in the case of the negative effect and not the positive effect (Guglielmo & Malle, 2010; Nanay, 2010; Sripada, 2009; Uttich & Lombrozo, 2010). Our findings suggest that good effects are seen as intended when no other intention information is present; by contrast, side-effect effect cases apparently provide extra information about the agent's intention towards one of the effects.

The second difference between the side-effect effect and our findings is that the former seems to be most robust when subjects are asked if the protagonist acted "intentionally"; by contrast, the effect is much weaker when subjects are asked if it was the agent's "intention" to bring about the good and bad effects (Knobe, 2004). As discussed in the introduction, our phenomenon primarily concerns goal-attribution, that is, acting with an intention (cf. Premack, 1990; Searle, 1983), and not acting "intentionally" or non-accidentally.

Our results generate a new set of questions about the good intention prior. For example, how much (and what quality) of countervailing information is required to override the prior? Where might this information come from? One source of evidence about an agent's intentions might be information about an agent's character. If people have background knowledge about an agent (e.g., that he is known to be bad or has a history of intending bad effects), then the information about the agent's character could shift their priors away from the default. If priors are built up from observing agents' good and bad intentions, then it seems likely that this prior knowledge and information about the characters of agents are in dynamic exchange with one another.[7] More research is needed to clarify this issue.

Conclusion

A recurring problem in the theory of moral cognition is to explain how individuals manage to determine the goals and intention structure of an action in the absence of clear or unambiguous evidence. Extending previous work on this topic (Mikhail, 2007, pp. 146–148; Mikhail, 2011, pp. 162–174), the studies presented in this paper supply the first empirically grounded solution to this problem by showing that a good intention prior appears to develop by 3 years of age and persists in adult cognition. This prior explains how children and adults can infer the goals of novel actions with more than one salient effect. The prior is conceptually similar to one element of the "presumption of innocence" one finds in both domestic criminal law and international human rights law (see, e.g., Article 11 of the Universal Declaration of Human Rights). Although this prior knowledge might be learned in early development, it also seems like a good candidate to be built into core cognition, a kind of cognitive constraint that enables morally relevant intention structures to be generated in infants and young children in the absence of sufficient information from the environment (cf. Spelke et al., 2013). Indeed, in explaining this presumption of good intentions, Mikhail (2011, pp. 172–173) points to a long philosophical tradition presupposing that humans possess innate moral knowledge, including the precept to "pursue good and avoid evil" (e.g., Hume, 1740/1978: 438; Aquinas, 1274/1988: 49). Without this rebuttable presumption, ordinary 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 many other basic social interactions would be rendered futile. That we normally assume good intentions and give others the benefit of the doubt binds our social world together.

[7] We thank an anonymous reviewer for bringing this point to our attention.

References

Aquinas, T. (1988). In P. Sigmund (Ed.), *St. Thomas Aquinas on politics and ethics*. New York, NY: Norton. (Original work published 1274)

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development, 2004,* 37–49. http://dx.doi.org/10.1002/cd.96

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113,* 329–349. http://dx.doi.org/10.1016/j.cognition.2009.07.005

Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2005). Bayesian models of human action understanding. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in Neural Information Processing Systems* (Vol. 18, pp. 99–106). Cambridge, MA: MIT Press.

Baron-Cohen, S. (Ed.). (1997). How to build a baby that can read minds: Cognitive mechanisms in mindreading. *The maladapted mind: Classic readings in evolutionary psychopathology* (pp. 207–239). East Sussex, UK: Psychology Press.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8,* 536–554.

Behne, T., Carpenter, M., & Tomasello, M. (2005). One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental Science, 8,* 492–499. http://dx.doi.org/10.1111/j.1467-7687.2005.00440.x

Biro, S., & Leslie, A. M. (2007). Infants' perception of goal-directed actions: Development through cue-based bootstrapping. *Developmental Science, 10,* 379–398. http://dx.doi.org/10.1111/j.1467-7687.2006.00544.x

Bratman, M. (1987). *Intention, plans, and practical reason*. United Kingdom: Cambridge University Press.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences, 17,* 363–366.

Csibra, G., Bıró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science, 27,* 111–133. http://dx.doi.org/10.1207/s15516709cog2701_4

Csibra, G., & Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental Science, 1,* 255–259. http://dx.doi.org/10.1111/1467-7687.00039

Csibra, G., & Gergely, G. (2007). 'Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica, 124,* 60–78. http://dx.doi.org/10.1016/j.actpsy.2006.09.007

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17,* 273–292. http://dx.doi.org/10.1177/1088868313495594

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127,* 6–21. http://dx.doi.org/10.1016/j.cognition.2012.11.008

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17,* 1082–1089. http://dx.doi.org/10.1111/j.1467-9280.2006.01834.x

Di Giorgio, E., Lunghi, M., Simion, F., & Vallortigara, G. (2017). Visual cues of motion that trigger animacy perception at birth: The case of self-propulsion. *Developmental Science, 20,* 1–12. http://dx.doi.org/10.1111/desc.12394

Doris, J. M., & Moral Psychology Research Group. (2010). *Moral psychology handbook*. http://dx.doi.org/10.1093/acprof:oso/9780199582143.001.0001

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2,* 493–501.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences, 7,* 287–292. http://dx.doi.org/10.1016/S1364-6613(03)00128-1

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56,* 165–193. http://dx.doi.org/10.1016/0010-0277(95)00661-H

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96,* 1029.

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York, NY: Penguin.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111,* 364–371. http://dx.doi.org/10.1016/j.cognition.2009.02.001

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin, 36,* 1635–1647. http://dx.doi.org/10.1177/0146167210386733

Hamlin, J. K. (2015). The infantile origins of our moral brains. *The moral brain: A multidisciplinary perspective*. Cambridge, MA: MIT Press.

Hamlin, J. K., & Baron, A. S. (2014). Agency attribution in infancy: Evidence for a negativity bias. *PLoS ONE, 9*(5), e96112. http://dx.doi.org/10.1371/journal.pone.0096112

Hamlin, J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science, 16,* 209–226. http://dx.doi.org/10.1111/desc.12017

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., . . . Hill, K. (2007). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *International Library of Critical Writings in Economics, 204,* 343.

Hume, D. (1978). *A treatise of human nature*. Oxford, UK: Clarendon Press. (Original work published 1740)

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134,* 193–209.

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition, 119,* 197–215. http://dx.doi.org/10.1016/j.cognition.2011.01.006

Killen, M., & Smetana, J. G. (2015). Origins and development of morality. In R. M. Lerner & M. E. Lamb (Eds.), *Handbook of child psychology and developmental science* (7th ed., Vol. 3, pp. 701–749). New York, NY: Wiley-Blackwell. http://dx.doi.org/10.1002/9781118963418.childpsy317

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63,* 190–194. http://dx.doi.org/10.1093/analys/63.3.190

Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis, 64,* 181–187. http://dx.doi.org/10.1093/analys/64.2.181

Knobe, J., & Mendlow, G. S. (2004). The good, the bad and the blameworthy: Understanding the role of evaluative reasoning in folk psychology. *Journal of Theoretical and Philosophical Psychology, 24,* 252–258. http://dx.doi.org/10.1037/h0091246

Leslie, A. M. (1991). The theory of mind impairment in autism: Evidence for a modular mechanism of development? In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading* (pp. 63–78). Cambridge, MA: Basil Blackwell.

Leslie, A. M. (1994). ToMM, ToBy, and agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511752902.006

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect. *Psychological Science, 17,* 421–427. http://dx.doi.org/10.1111/j.1467-9280.2006.01722.x

Leslie, A. M., Mallon, R., & DiCorcia, J. A. (2006). Transgressors, victims, and cry babies: Is basic moral judgment spared in autism? *Social Neuroscience, 1,* 270–283. http://dx.doi.org/10.1080/17470910600992197

MacHery, E. (2008). Understanding the folk concept of intentional action: Philosophical and experimental issues. *Mind & Language, 23,* 101–121. http://dx.doi.org/10.1111/j.1468-0017.2007.00336.x

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25,* 147–186.

Margoni, F., & Surian, L. (2017). Children's intention-based moral judgments of helping agents. *Cognitive Development, 41,* 46–64.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Meltzoff, A. N. (2005). Imitation and other minds: The "like me" hypothesis. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation: From neuroscience to social science* (Vol. 2, pp. 55–77). Cambridge, MA: MIT Press.

Meltzoff, A. N. (2007). The "like me" framework for recognizing and becoming an intentional agent. *Acta Psychologica, 124,* 26–43. http://dx.doi.org/10.1016/j.actpsy.2006.09.005

Mikhail, J. M. (2000). *Rawls' linguistic analogy: A study of the "generative grammar" model of moral theory described by John Rawls in "a theory of justice"* (Doctoral dissertation). Retrieved from Social Science Research Network.

Mikhail, J. (2002). *Aspects of the theory of moral cognition: Investigating intuitive knowledge of the prohibition of intentional battery and the principle of double effect* (Georgetown University Law Center Public Law & Legal Theory Working Paper No. 762385). Retrieved from http://ssrn.com/abstract=762385

Mikhail, J. (2005). Moral heuristics or moral competence? Reflections on Sunstein. *Behavioral and Brain Sciences, 28,* 557–558. http://dx.doi.org/10.1017/S0140525X05380095

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11,* 143–152. http://dx.doi.org/10.1016/j.tics.2006.12.007

Mikhail, J. (2009). Moral grammar and intuitive jurisprudence: A formal model of unconscious moral and legal knowledge. *Psychology of Learning and Motivation, 50,* 27–100.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511780578

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General, 138,* 535–545. http://dx.doi.org/10.1037/a0016796

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16,* 406–419. http://dx.doi.org/10.1037/a0024377

Nanay, B. (2010). Morality of modality? What does the attribution of intentionality depend on? *Canadian Journal of Philosophy, 40,* 25–40. http://dx.doi.org/10.1353/cjp.0.0087

Pellizzoni, S., Siegal, M., & Surian, L. (2010). The contact principle and utilitarian moral judgments in young children. *Developmental Science, 13,* 265–270. http://dx.doi.org/10.1111/j.1467-7687.2009.00851.x

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition, 36,* 1–16. http://dx.doi.org/10.1016/0010-0277(90)90051-K

Rakoczy, H., Behne, T., Clüver, A., Dallmann, S., Weidner, S., & Waldmann, M. R. (2015). The Side-effect effect in children is robust and not specific to the moral status of action effects. *PLoS ONE, 10*(7), e0132933. http://dx.doi.org/10.1371/journal.pone.0132933

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21,* 301–308. http://dx.doi.org/10.3758/s13423-014-0595-4

Saunders, K. (2014). *Investigating the psychological foundations of moral judgment* (Doctoral dissertation). Rutgers University-Graduate School-New Brunswick, New Brunswick, NJ.

Scholl, B. J. (2005). Innateness and (Bayesian) visual perception. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Structure and contents* (pp. 34–52). New York, NY: Oxford University Press.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4,* 299–309. http://dx.doi.org/10.1016/S1364-6613(00)01506-0

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language, 27,* 135–153. http://dx.doi.org/10.1111/j.1468-0017.2012.01438.x

Scott, R. M., & Baillargeon, R. (2013). Do infants really expect agents to act efficiently? A critical test of the rationality principle. *Psychological Science, 24,* 466–474. http://dx.doi.org/10.1177/0956797612457395

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139173452

Sommerville, J. A., & Crane, C. C. (2009). Ten-month-old infants use prior information to identify an actor's goal. *Developmental Science, 12,* 314–325. http://dx.doi.org/10.1111/j.1467-7687.2008.00787.x

Sommerville, J. A., & Woodward, A. L. (2005). Pulling out the intentional structure of action: The relation between action processing and action production in infancy. *Cognition, 95,* 1–30. http://dx.doi.org/10.1016/j.cognition.2003.12.004

Spelke, E. S., Bernier, E. P., & Skerry, A. E. (2013). Core social cognition. In M. Banaji & S. Gelman (Eds.), *Navigating the social world. What infants, children, and other species can teach us* (pp. 11–16). New York, NY: Oxford University Press.

Sripada, C. S. (2009). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies, 151,* 159–176. http://dx.doi.org/10.1007/s11098-009-9423-5

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception, 29,* 943–951. http://dx.doi.org/10.1068/p3101

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition, 116,* 87–100. http://dx.doi.org/10.1016/j.cognition.2010.04.003

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779–804. http://dx.doi.org/10.3758/BF03194105

Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York, NY: Spring Science + Business Media.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 362–389). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780199734689.013.0019

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science, 18,* 247–253. http://dx.doi.org/10.1111/j.1467-9280.2007.01884.x

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69,* 1–34. http://dx.doi.org/10.1016/S0010-0277(98)00058-4

Woodward, A. L. (2013). Infant foundations of intentional understanding. In M. Banaji & S. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 75–80). New York, NY: Oxford University Press.

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science, 11,* 73–77. http://dx.doi.org/10.1111/1467-9280.00218

Woodward, A. L., Sommerville, J. A., Gerson, S., Henderson, A. M., & Buresh, J. (2009). The emergence of intention attribution in infancy. *Psychology of Learning and Motivation, 51,* 187–222. http://dx.doi.org/10.1016/S0079-7421(09)51006-7

Woodward, A. L., Sommerville, J. A., & Guajardo, J. J. (2001). How infants make sense of intentional action. In B. Malle, L. Moses, & D. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 149–169). Cambridge, MA: MIT Press.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 104,* 8235–8240. http://dx.doi.org/10.1073/pnas.0701408104

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120,* 202–214. http://dx.doi.org/10.1016/j.cognition.2011.04.005

Appendix

Stimuli for Experiments 1–3

Stimuli for Experiment 1: Adult Subjects

Uninformative Condition

Hank is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. There are five men walking across the tracks and 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.

Informative condition good. Joe is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. There are five men walking across the tracks and 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.

Informative Condition Bad. Mark is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. There are five men walking across the tracks and 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)

Why do you think Hank/Joe/Mark threw the switch? (Options: To save the five men on the main track/To kill the one man on the side track.)

(*Appendix continues*)

Stimuli for Experiment 2: Adult Subjects

Uninformative Condition

Hank is taking his daily walk near the train tracks when he notices that an empty runaway train is approaching. The train is rushing toward a man, who is walking across the tracks. It is moving so fast that he will not be able to get off the tracks 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 man. There are five men on the side track. Hank can throw the switch, killing the five men; or he can refrain from doing this, letting the man on the main track die. Hank throws the switch.

Informative condition good. Joe is taking his daily walk near the train tracks when he notices that an empty runaway train is approaching. The train is rushing toward a man, who is walking across the tracks. It is moving so fast that he will not be able to get off the tracks 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 man. There are five men on the side track. Hank can throw the switch, killing the five men; or he can refrain from doing this, letting the man on the main track die. Joe then recognizes that the man is someone who he deeply cares about. Joe thinks to himself, “This is my chance to save that man.” Joe throws the switch.

Informative Condition Bad. Mark is taking his daily walk near the train tracks when he notices that an empty runaway train is approaching. The train is rushing toward a man, who is walking across the tracks. It is moving so fast that he will not be able to get off the tracks 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 man. There are five men on the side track. Mark can throw the switch, killing the five men; or he can refrain from doing this, letting the man on the main track die. Mark then recognizes that the men on the side track are people who he hates with a passion. Mark thinks to himself, “This is my chance to kill those bastards.” Mark throws the switch.

Test questions for Experiment 2. Is it morally permissible for Hank/Joe/Mark to throw the switch? (Options: Yes/No)

Why do you think Hank/Joe/Mark threw the switch? (Options: To save the one man on the main track/To kill the five men on the side track.)

Stimuli for Experiment 3: Preschool Subjects

These stories were accompanied by animations shown to the subjects.

Uninformative 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 cannot 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 do not 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 do not remember, help them.

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

(*Appendix continues*)

Exclusion criteria. What did Sally do? [Answer: Put up the gate.]

What did the squirrel do? [Answer: Eat the one kid's cookie.]

How did that kid feel? [Answer: Sad.]

Were those kids sad? [Answer: Not sad/happy.]

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?

Informative condition good. 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 cannot 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 do not 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 do not remember, help them. If correct response, say "That's right."

Exclusion criteria. Does Sally like these kids? (the five) [Answer: yes.]

How does Sally want to make these kids feel? [Answer: Not sad/happy.]

What did Sally do? [Answer: Put up the gate.]

What did the squirrel do? [Answer: Eat the one kid's cookie.]

How did that kid feel? [Answer: Sad.]

Were those kids sad? [Answer: Not sad/happy.]

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?

Informative Condition Bad. 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.

(*Appendix continues*)

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. 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 cannot 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 do not 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 do not remember, help them.

If correct response, say "That's right."

Exclusion criteria. Does Sally like this kid? (the one) [Answer: No.]

How does Sally want to make this kid feel? [Answer: Sad/bad.]

What did Sally do? [Answer: Put up the gate.]

What did the squirrel do? [Answer: Eat the one kid's cookie.]

How did that kid feel? [Answer: Sad.]

Were those kids sad? [Answer: Not sad/happy.]

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?

Pink scale screening. Children were shown the scale pictured in Figure A1.

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 cannot think of something good at all (or cannot think of something really good or a little good) offer a suggestion such as "eating an apple" or "helping your teacher" or "playing outside."

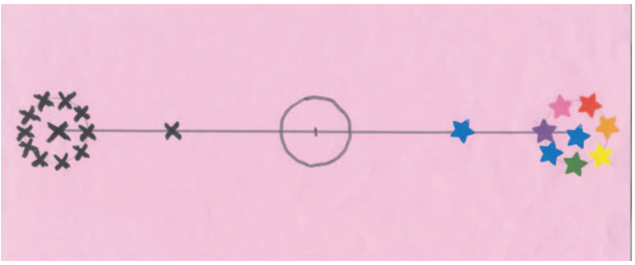

Figure A1. Likert scale used by children to rate the action of the agent in Experiment 3. See the online article for the color version of this figure.

(*Appendix continues*)

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?

Received August 9, 2017
Revision received April 5, 2018
Accepted April 15, 2018 ■