Cognition 127 (2013) 6–21

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

The development of intent-based moral judgment

Fiery Cushman [a,*], Rachel Sheketoff [b], Sophie Wharton [c], Susan Carey [b]

[a] Department of Cognitive, Linguistic and Psychological Sciences, Brown University, 89 Waterman St., Providence, RI 02912, United States
[b] Department of Psychology, Harvard University, 33 Kirkland St., Cambridge MA 02138, United States
[c] Department of Psychology, New York University, 6 Washington Place, NY 10003, United States

ARTICLE INFO

Article history:
Received 21 February 2012
Revised 13 November 2012
Accepted 15 November 2012
Available online 12 January 2013

Keywords:
Morality
Development
Theory of mind
Moral luck

ABSTRACT

Between the ages of 4 and 8 children increasingly make moral judgments on the basis of an actor's intent, as opposed to the outcome that the actor brings about. Does this reflect a reorganization of concepts in the moral domain, or simply the development of capacities outside the moral domain such as theory of mind and executive function? Motivated by the past evidence that adults rely partially on outcome-based judgment for judgments of deserved punishment, but not for judgments of moral wrongness, we explore the same categories of judgment in young children. We find that intent-based judgments emerge first in children's assessments of naughtiness and that this subsequently constrains their judgments of deserved punishment. We also find that this developmental trajectory differs for judgments of accidental harm (a bad outcome with benign intent) and judgments of attempted harm (a benign outcome with bad intent). Our findings support a two process model derived from studies of adults: a mental-state based process of judging wrongness constrains an outcome-based process of assigning punishment. The emergence of this two-process architecture in childhood suggests that the developmental shift from outcome- to intent-based judgment involves a conceptual reorganization within the moral domain.

1. The development of intent-based moral judgment

In many circumstances when preschoolers judge the moral valence of an act, they focus on its outcome, largely ignoring the actor's beliefs and intentions. Between ages 4 and 10, they increasingly take mental states into account when judging whether an act is wrong, or how severely it should be punished. Piaget (1965/1932) famously demonstrated that, for instance, young children consider it morally worse to accidentally make a large ink stain than to intentionally make a very small one (indicating a focus on the severity of the outcome), whereas older children make the opposite judgment (indicating a focus on malicious intent[1]). Since that seminal investigation the developmental shift from outcome- to intent-based moral judgment has been extensively documented (e.g. Armsby, 1971; Baird & Astington, 2004; Costanzo, Coie, Grumet, & Farnill, 1973; Hebble, 1971; Imamoglu, 1975; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; Shultz, Wright, & Schleifer, 1986; Yuill & Perner, 1988; Zelazo, Helwig, & Lau, 1996).

Yet, the psychological basis of the "outcome-to-intent shift" is still poorly understood—so much so, in fact, that it is uncertain whether it reflects genuine conceptual

[1] Although this developmental shift is typically described as decreasing "outcome" focus and increasing "intent" focus, children's moral evaluations of "intent" encompass of other mental states such as foresight (Yuill & Perner, 1988), and children's evaluations of "outcome" encompass a causal relation between agent and outcome, and not the mere occurrence of an outcome by any cause (Fincham and Jaspers, 1979).

* Corresponding author.
E-mail address: Fiery_Cushman@brown.edu (F. Cushman).

0010-0277/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2012.11.008

change within the moral domain. Piaget (1965/1932) argued for a fundamental reorganization of moral concepts during childhood, a position later elaborated by Kohlberg (1969). These "stage theory" accounts of the outcome-to-intent shift posited that (1) an early outcome-alone based concept of morality is fully replaced by a later intent-based concept of morality between ages 6 and 10, and (2) that this reorganization of concepts the moral domain is enabled by the child's acquisition of domain-general capacities for abstract thought, and a domain-general shift away from egocentric thought.

Both of these posits have been forcefully challenged. Several studies show that young children's moral judgments are sensitive to intent in children as young as 3 years old when stimuli are carefully controlled to remove confounding factors (Armsby, 1971; Farnill, 1974; Yuill & Perner, 1988). Indeed, many results indicate that Piaget and Kohlberg underestimated the moral concepts of young children, finding instead many continuities in moral reasoning over development. For example, 3-year-olds distinguish between moral and conventional restrictions on action (Smetana, 1981), and even young babies have negative reactions to agents who harm others (Hamlin, Wynn, & Bloom, 2007). There is also some evidence for sensitivity to intent in the moral judgment preschoolers (Nobes, Panagiotaki, & Pawson, 2009; Yuill & Perner, 1988) and in the evaluative judgments of infants (Hamlin, Ullman, Tenenbaum, Goodman, & Baker, submitted for publication). Conversely, other studies show that adults' moral judgments maintain some sensitivity to outcome, sometimes in the absence of any intent to harm (Berg-Cross, 1975; Cushman, 2008; Cushman, Dreber, Wang, & Costa, 2009; Gino, Moore, & Bazerman, 2008; Gino, Shu, & Bazerman, 2010; Mazzocco, Alicke, & Davis, 2004). While a developmental *increase* in intent-based moral judgment is beyond dispute, these data contradict a theory of wholesale conceptual replacement. Moreover, Piaget's model of a broad, domain-general shift from concrete to abstract thought fails across diverse case-studies of domain-specific conceptual changes in childhood (Carey, 1985; Gelman & Baillargeon, 1983). And, Kohlberg's emphasis on the child's controlled application of explicit moral theories largely ignores the role of automatic processes of moral judgment (Cushman, Young, & Hauser, 2006; Greene, 2008; Haidt, 2001; Shweder & Haidt, 1993).

Perhaps because of these challenges, recent discussions of the outcome-to-intent shift take a markedly different approach. They side-step the issue of whether the outcome-to-intent shift reflects a reorganization of moral concepts by linking it to developmental changes outside the moral domain. For instance, two studies correlate intent-based moral judgment with developmental changes in theory of mind (Chandler, Sokol, & Hallett, 2001; Killen et al., 2011). The authors suggest that intent-based moral judgment cannot be expressed until the child possesses the capacity to represent others' mental states in sufficiently rich detail. Another approach suggests that the outcome-to-intent shift reflects changes in executive function that enable the child to integrate information about intent and outcomes during the process of moral judgment (Zelazo et al., 1996).

These studies make important strides in articulating developmental prerequisites for the outcome-to-intent shift, but they still leave several questions unanswered about the nature of the shift itself. Does it entail any conceptual reorganization within the moral domain, or are changes in theory of mind and executive function sufficient to allow the expression of a latent capacity for intent-based moral judgments without conceptual reorganization? And, to what extent are outcome- and intent-based moral judgments the product of explicit moral theories, as opposed to automatic processes that give rise to moral intuition?

Answering these questions requires us to individuate the psychological mechanisms that contribute to outcome-based and mental-state-based moral judgment. Recent studies of adult moral judgment take up this task, and thus they offer a useful guide for developmental research. These studies suggest that even in adulthood there are two distinct processes that enter into moral judgments: one that analyzes the causes of harmful outcomes and one that analyzes the intentions and knowledge states of the actors. They show that these two distinct processes sometimes conflict and enter somewhat independently into different moral judgments (e.g., of punishability and wrongness). Here we draw on this literature to offer a new perspective on one of the oldest findings in moral development as well as to assess whether the outcome-to-intent shift reflects a conceptual reorganization within the moral domain.

1.1. A two-process theory: evidence from adults

Research in moral psychology has long attempted to characterize the processes by which causal and mental state attributions contribute to our moral judgments (e.g. Cushman et al., 2006; Darley & Shultz, 1990; Greene et al., 2009; Heider, 1958; Mikhail, 2002; Weiner, 1995). Existing models of moral judgment typically posit a single process that integrates information about an agent's causation of harm and about the agent's mental states in order to deliver an ultimate moral verdict (e.g. Darley & Shultz, 1990; Heider, 1958; Mikhail, 2007; Weiner, 1995). For instance, several models propose that actions are deemed wrong or punishable when both factors are present: a person causes harm and the harm was committed with a culpable mental state such as intent or foresight (e.g. Darley & Shultz, 1990; Heider, 1958; Weiner, 1995; Zelazo et al., 1996).

Recent evidence indicates quite a different arrangement, however: a competitive interaction between independent processes of moral judgment, one that depends on causal attributions for harmful outcomes and another that depends on mental state information (Cushman, 2008; Young, Cushman, Hauser, & Saxe, 2007). A unique prediction of a two-process model is competition and conflict between processes. On a single-process model, causal and mental-state information have no opportunity to generate moral conflict—there is no moral judgment until both sources of information have been considered and integrated. On a two-process model, however, instances where harm is caused but not intended (an accident) or where

harm is intended but not caused (an attempt) could lead to conflict between two contrasting moral judgments, and thus competition for the ultimate "verdict".

This model may explain the philosophical dilemma of moral luck (Nagel, 1979; Williams, 1981). Simply put, the problem of moral luck is that chance circumstances contribute to moral evaluation. For instance, if one reckless driver hits a person while another hits a tree, the former receives greater punishment and blame than the latter just because of the "luck" of what their out-of-control vehicle happened to strike. Cases like these may generate philosophical controversy because the mind furnishes two solutions to the same problem: one predicated on mental states and controllable action that rejects the role of outcome-based luck, and another predicated on causal responsibility for a harmful outcome that facilitates the effect of luck (Cushman & Greene, 2012).

1.2. Blame blocking

One source of evidence for competition between processes comes from a phenomenon termed blame blocking (Cushman, 2008). Specifically, people assign less blame and punishment to a person's attempted crime (e.g. shooting at a victim but missing) if the intended victim happens to be killed by some alternative mechanism (e.g. the victim happens to be struck by lightning) than if they are not killed at all. Apparently, the presence of a harmful outcome (the lightning strike) triggers a process of causal attribution that points away from the attempted harmdoer (the shooter), leading to a reduction in judgments of deserved punishment and blame. By contrast, in the absence of any harmful outcome no analysis of causal responsibility is triggered, and thus the analysis of the attempted harmdoer's malicious mental state proceeds unabated. This effect is difficult to explain on a single-process model of moral judgment, in which the assessment of causal responsibility should not competitively block the consideration of malicious mental states.

1.3. Accidental vs. attempted harms

Further evidence of conflict between processes derives from research using functional neuroimaging (Young et al., 2007). When people judge cases of accidental harm (+caused harm, −intent) they exhibit increased activity in brain regions associated with cognitive conflict and top-down control, relative to cases of intentional harm (+caused harm, +intent). Notably, however, no equivalent signature of cognitive conflict was observed in cases of attempted, but failed, harm (−caused harm, +intent) relative to benign cases (+caused positive outcome, −intent). This result suggests that there is something especially difficult about exculpating accidental harm-doers. Just as with blame blocking, it appears that the "causal process" of moral judgment is triggered by the presence of a harmful outcome, but remains silent in the absence of a harmful outcome. Thus, when a harmful outcome occurs but there is no intent (an accident), conflict arises between the causal process and the mental state process. But when no harmful outcome occurs in the presence of malicious intent (an attempt), conflict does not arise; the causal process remains silent, and the mental state process operates unabated.

Convergent evidence from several additional studies corroborates the claim that cases of accidental harm engage a distinct set of mechanisms compared with cases of attempted harm. First, high-functioning individuals with autism spectrum disorder tend to judge accidental harms morally worse compared with a control group, but tend to judge attempted harms no differently (Moran et al., 2011). Second, the disruption of the right temporoparietal junction (rTPJ) via transcranial magnetic stimulation consistently causes individuals to judge attempted harms morally worse than controls do, but has an inconsistent effect on the judgment of accidental harms (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). (Activity in rTPJ has been frequently associated with reasoning about others' mental states generally, and about their false beliefs in particular). Third, individuals with damage to the ventromedial prefrontal cortex (VMPFC) tend to judge attempted harms less morally wrong compared with a control group, but tend to judge accidental harms no differently (Young, Bechara et al., 2010). Although a consistent mapping between psychological process and neural substrate has yet to be fully developed, and the state of research is still in flux, in each of these cases, atypical neural processing leads to lesser reliance on mental states during moral judgment selectively: for either accidents or for attempts, but not both. This is consistent with the claim that unique psychological processes are necessary to judge accidental harms because of the conflict between mental-state based and outcome based processes.

1.4. Judgments of wrongness versus deserved punishment

A final piece of evidence for two distinct processes of moral judgments derives from the fact that judgments of moral wrongness rely relatively more on an agents' mental states, whereas judgments of deserved punishment rely relatively more on an agents causal responsibility for harm. For instance, consider two nannies who each thoughtlessly leave infants in a locked car on a hot day while picking up groceries. The first nanny's car happens to leave its vents open, and therefore the infant survives unharmed. But, the second nanny's car happens to close its vents automatically, and so the infant dies of suffocation and heat exposure. People would tend to judge the two actions equally "morally wrong" based on their identical mental states and behavior (Cushman, 2008). They would also tend to judge the two nannies roughly equally in terms of moral character (Cushman, unpublished data). But, they would tend to assign much more punishment to the nanny who kills and less punishment to the nanny who does not because of the very different outcomes that their behavior causes (Cushman, 2008; Cushman et al., 2009). These results suggest that for adults, punishability is partially distinct from moral wrongness, and that the two hypothesized processes underlying moral judgment participate differentially in judgments of each.

1.5. Synthesizing stage theories and two-process theory

The evidence for a two-process theory dovetails with several core features of Piaget and Kohlberg's stage theories of moral development. Piaget interpreted the outcome-to-intent shift as part of a broader change in the young child's moral concepts, which in turn were part of a domain-general shift from concrete to abstract thought, and away from egocentric thought. His theory of moral development centered on two broad stages: An earlier "heteronomous" stage in which the child conceives of morality in terms of unidirectional authority relations (e.g., "It's wrong to steal because my daddy says so"), and a later "autonomous" stage in which the child conceives of morality in terms of mutually agreed-upon standards of conduct developed between peers (e.g., "It's wrong to steal because I don't want people to steal from me"). He theorized that heteronomous-stage children make moral judgments on the basis of outcome because the salient, overt marker of a "wrongful action" is punishment, and the salient, overt cause of punishment is a bad outcome. Meanwhile, autonomous-stage children tend to make moral judgments on the basis of intent as they acquire a concept of moral obligation based on the coordination of interests between peers, a concept that is more abstract in nature, and also less egocentric. Piaget writes, "It is when the child is accustomed to act from the point of view of those around him, when he tries to please rather than to obey, that he will judge in terms of intentions. So that taking intentions into account presupposes cooperation and mutual respect" (Piaget, 1965/1932, p. 137). In other words, the requirement of perspective taking inherent to a morality of reciprocity refocuses the child's judgment away from the consequence of what others do, and instead towards the way that they choose their actions, an act of mental-state assessment.

This position was later echoed by Kohlberg (1969), according to whom outcome-based moral judgment accompanied an "obedience and punishment orientation" in which "moral value resides in external, quasi-physical happenings [and] bad acts", whereas mental-state-based moral judgment accompanied an "orientation to approval and pleasing and helping others" in which "moral value resides in performing good or right roles".

There is an evident homology between the moral stages posited by Piaget and Kohlberg and the two process of moral judgment we posited above. One set of judgments assigns punishment on the basis of causal responsibility for harm, while another set of judgments assesses the wrongness of a moral action on the basis of the mental states that determined that action. But whereas Piaget and Kohlberg viewed these as explicit conceptual systems operating uniquely within distinct periods of child development, the two process model suggests that they operate in parallel among adults and is agnostic about their status as explicit conceptual systems versus automatic processes of moral judgment.

The present experiment tests a proposed synthesis of these two theories. We suggest that an early system of moral judgment yields judgments of punishment on the basis of causal responsibility (i.e., outcomes). And, initially, the processes that assign moral blame do not differentiate between judgments of wrongness and judgments of punishment (perhaps, as Piaget and Kohlberg suggested, because what is punishable is a major source of evidence as to what is wrong in the conceptual system of a preschool aged child). During the outcome-to-intent shift, the child acquires a new concept of moral wrongness that is grounded in the assessment of action and the mental states that give rise to action. At this point two related changes occur in their moral judgments: those judgments show increasing sensitivity to intent, and judgments of wrongness become differentiated from judgments of punishment. Now what is punishable follows, at least in part, from what is morally wrong. Critically, it is the emergence of an intent-based concept of moral wrongness that constrains judgments of punishment, such that they become increasingly reliant on intent as well. We refer to this as the "constraint hypothesis" (Fig. 1, top panel).

This constraint relationship between intent-based wrongness judgments and intent-based punishment judgments cannot be explained by the development of capacities outside the moral domain, such as theory of mind (Chandler et al., 2001; Killen et al., 2011) and executive function (Zelazo et al., 1996). Both a sophisticated theory of mind and executive control are presumably necessary for the expression of intent-based moral judgment, and development in these domains may help to explain its emergence. There is no special connection, however, between theory of mind or executive function, and judgments of wrongness versus punishment. Whatever role developmental processes outside the moral domain may play, the constraint hypothesis posits that the outcome-to-intent shift *also* entails a specific and fundamental reorganization of concepts within the moral domain.

Contrasting with the constraint hypothesis is an alternative that we call the parallel hypothesis (Fig. 1, bottom panel). The parallel hypothesis holds that intent-based moral judgment emerges simultaneously for different categories of judgment (wrongness, punishment, etc.). The parallel hypothesis is the default prediction if the outcome-to-intent shift is fully driven by changes outside of the moral domain (e.g. to theory of mind, or executive

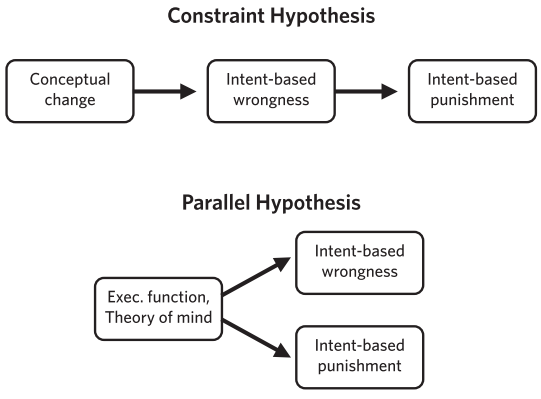

Fig. 1. The constraint and parallel hypotheses.

function) and without any conceptual reorganization within the moral domain.

In order to test these competing hypotheses we presented 4- through 8-year-old children with two stories, one involving an attempted, but failed, harm (e.g. Cliff tosses a ball at Mom's mirror but instead it lands in the box where it belongs) and one involving an accidental harm (e.g. Jack tosses a ball at the box where it belongs but instead it hits and breaks Mom's mirror). After each story participants were asked to judge both whether the agent was "bad" and "naughty", and also whether the agent deserved to be punished. Our analyses focus on two dimensions of this design: judgments of punishability versus naughtiness, and judgments of accidental harms versus attempted harms.

At a broad level, if the developmental shift from outcome-based to intent-based moral judgment reflects the emergence of the adult two-process system, then it should coincide with the differentiation of wrongness from punishment judgments in terms of their relative dependency on outcome versus intent.

Our study affords two precise and independent tests of the constraint hypothesis. The first tests a developmental-level prediction of the constraint hypothesis using mediation analysis (Baron & Kenny, 1986). Specifically, it asks whether the effect of age on intent-based punishment judgment for accidental harms is statistically mediated by the effect of age on intent-based wrongness judgments, as predicted by the constraint hypothesis. In plain terms, do children start making intent-based wrongness judgments first, and *as a consequence* later begin making intent-based punishment judgments?

The second leverages order effects to test a processing-level prediction of the constraint hypothesis. When a naughtiness judgment directly precedes a punishment judgment it should exert a strong influence on that punishment judgment due to its constraining role. By contrast, when a punishment judgment directly precedes a naughtiness judgment it is not predicted to exert a strong influence on that naughtiness judgment because punishability does not function as a constraint on naughtiness. In other words, the constraint hypothesis predicts a unidirectional order effect: punishment judgments are influenced by prior naughtiness judgments, but not the other way around. Our experiment explores these predictions.

2. Methods

2.1. Participants

Participants were recruited in the Discovery Center at the Museum of Science in Boston, a hands-on learning environment for young children visiting the Museum. Researchers approached either the participant or participant's parent and proposed participating a short experiment. Experiments were conducted at a table in the corner of the Discovery Center. All children were tested in view of their parent or guardian, who provided signed consent. Participants were 293 4- to 8-year-old children (age 4, $n = 61$, age 5, $n = 65$; age 6 $n = 64$, age 7, $n = 55$, age 8, $n = 48$), 141 female. An additional three children (ages 4, 5 and 8) participated in the study, but were not included in the analyses because they failed to provide a single appropriate response. An additional 20 4-year-olds and 16 5-year-olds participated in a preliminary method check study. We did not record information on ethnicity or SES, but it is likely that families visiting the Museum of Science in Boston were more highly educated and more wealthy than the general population.

2.2. Stimuli: main study

Participants were presented with two stories drawn from one of four possible story contexts. Stories were presented one-at-a-time, and judgments were solicited and recorded for the first story before presenting the second story. One story involved an accidental harm, and the other story involved an attempted harm. Stories were read out loud by the experimenter and accompanied with illustrations. Synopses of the stories follow (the full text is provided in Supplementary Online Material):

"*Apple*" *context*: One boy accidentally steals an apple after it rolls into his shopping basket when he isn't looking. Another boy attempts to steal an apple, but it rolls out of his shopping basket when he isn't looking.

"*Ball*" *context*: One boy accidentally breaks a mirror when he throws a ball towards the bin where it belongs. Another boy attempts to break the mirror with the ball, but it lands in the bin where it belongs.

"*Push*" *context*: One boy is running when he trips on a rock and accidentally pushes somebody over. Another boy attempts to push somebody over when he trips on a rock and misses.

"*Paint*" *context*: One girl accidentally spills paint on the floor when a paint can slips out of her hand. Another girl attempts to spill paint on the floor, but the top is securely on the can, so none spills.

After each story participants answered two comprehension probes: (1) "Did [the character] want [the relevant outcome to occur]?", and (2) "Did [the character] actually [produce the relevant outcome]?". A positive side-effect of these comprehension probes was to focus children's attention on the features of the story most relevant to our investigation. Then, participants answered two test questions: (1) "Should [the character] be punished?", and (2) "Is [the character] a bad, naughty [boy/girl]?". Participants were prompted for a simple yes or no response to each question, rather than a continuous rating on a Likert scale as has been used on some past research. We chose to solicit dichotomous responses in order to facilitate comparison between the punishment and naughtiness dependent measures; had we used Likert scales, differences in judgment across the two dependent measures might reflect the interpretation or use of the scale rather than the underlying judgment process.

Three parameters were randomly assigned to each participant: the order of presentation of the accidental versus attempted versions of the story, the order of presentation of the "wanted to" versus "actually did" fact checks, and the order of presentation of the "punishment" versus "naughtiness" test questions. Additionally, different names

were used for the protagonist across the two stories presented to each child, along with slightly different physical appearances in the accompanying graphics, and these were counterbalanced between subjects.

2.3. Stimuli: preliminary method check study

To ensure that 4- and 5-year-olds could understand these stories and our moral probes, we created two morally unambiguous versions of a single scenario. In one, a case of intentional harm, a boy throws a ball at a mirror intending to break it and succeeds in breaking it. In the contrasting story, intentional good behavior, a boy throws a ball in the bin where it belongs and succeeds in putting it away. Each participant heard both stories, in a counterbalanced order. After hearing the story, the child was given the same comprehension and moral judgment probes as in the main study. *A priori*, our comprehension concern was greatest for the term "punishment"—we presume that most 4 year olds comprehend what it means to call someone a "bad, naughty" boy or girl. Consequently, every participant first judged deserved punishment, and then judged the naughtiness of the character.

2.4. Procedure

The experimenter told children that they would listen to a story and that they would be asked questions at the end. Children were told to listen carefully so that the questions would be easy. The experimenter then read the first story. Participants answered the fact checks and test questions associated with this story. Participants who expressed uncertainty were urged, "just do your best", and offered the opportunity to hear the story again. After completing these questions, participants were read the remaining story drawn from the same scenario context, and then responded to the same four questions in the same order.

3. Results and discussion

3.1. Preliminary method check

The morally unambiguous method check scenarios tested 4 year-olds' comprehension of the task and our moral judgment probes, which used the unambiguous versions of the ball-mirror scenario. Of 40 trials across 20 participants, 12 trials were excluded from analysis due to failure on fact checks. Eight of these failures involved denying that a character "wanted to break the mirror" when in fact the story specified that the character did want to break it. Apparently, young children have some difficulty accepting that people sometimes want to cause harm.

So long as they had passed the comprehension probes, the children in the method check study were more likely to assign punishment to intentionally harmful (70%) than intentionally harmless behavior (17%, Fisher Exact test $p = .01$, $N = 28$) and rated intentionally harmful actors more naughty (100%) than intentionally harmless actors (17%, Fisher Exact test $p < .001$, $N = 28$). This indicates that 4 year-old participants were capable of understanding our moral judgment probes.

However, the fact that a quarter of the 4-year-olds erred on at least one of the two comprehension probes indicates that they were at the limit of their capacity when tracking whether children were acting accidentally or on purpose. Accordingly, in the main study we include data only on those trials in which children succeed at the comprehension probes. In contrast, the 5-year-olds exhibited perfect performance on comprehension probes for the same scenarios, and 14/16 successfully distinguished the two scenarios in terms of both naughtiness and deserved punishment.

3.2. Main study

We excluded responses to any story for which the participant provided an incorrect answer to either of the two associated fact checks (10% of responses: age 4 = 28%, age 5 = 6%; age 6 = 10%, age 7 = 4%, age 8 = 2%). Again, such errors occurred mainly among 4-year-olds. Additionally, four participants are excluded from analysis by logistic regression because their age was recorded with precision to the year rather than to the day.

3.3. The outcome-to-intent shift

Increasing sensitivity to intent with development would be seen in an age dependent increase of moral condemnation of attempted harms, and decrease of moral condemnation of accidental harms. As can seen in Fig. 2, both effects were found. Collapsing over punishment and naughtiness judgments, and over the first and second scenarios, analysis by logistic regression revealed an age-related increase in condemnation of judgments of attempted (but failed) harms (Fig. 2; $\beta = .15$, $t = 2.45$, $p = .015$), and an age-related decrease in condemnation of accidental harms ($\beta = .25$, $t = 4.05$, $p < .001$). Comparing these results with our preliminary method check, we find that 4-year-olds judge attempted, but failed harms,

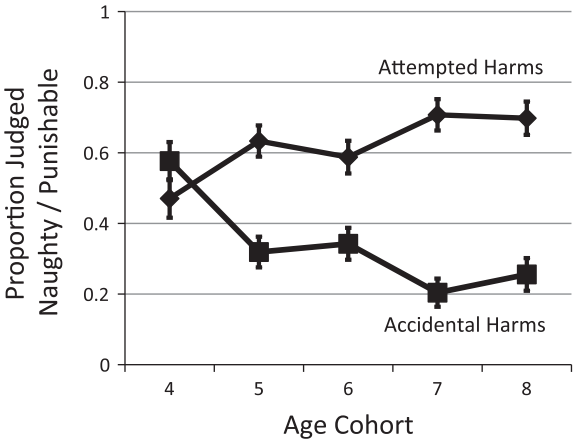

Fig. 2. Proportion of children who judged the character naughty or punishable collapsing ages 4 through 8, analyzing each child's response to all trials.

naughty and punishable (45%) more than the benign scenarios of the method check study (17%), and judge accidental harms naughty and punishable (59%) less than intentional harm scenarios of the method check study (90%). Thus, even 4-year-olds are sensitive both to outcome and intent in their moral judgments, and the relative weighting of these two factors changes with age. We again replicate the robust finding first observed by Piaget that young children rely relatively more on outcomes in their moral judgments, with increasing sensitivity to intent over the ages of 4 through 8.

As is evident in Fig. 2, however, the magnitude of the outcome-to-intent shift is larger for judgments of accidental harm (a shift of about 40% points) compared with judgments of attempted harm (a shift of about 20% points). Further inspection of the data showed that the effect of age on the judgment of accidents versus attempts interacts strongly with the order in which the two stories (accident and attempt) are presented to the child. Presumably the most accurate representation of children's spontaneous moral judgments comes from responses to the first story presented, whereas patterns of responses unique to the second story presented are likely to reflect preservation, explicit comparisons between the stories, etc. Analyzing responses to the first story only (see Fig. 3, top panel), and collapsing across judgments of punishment and wrongness, there is a strong negative correlation between age and the condemnation of accidental harms $r = -.40$, $p < .001$, but no correlation between age and condemnation of attempted harms $r = .03$, $p = .34$. This result converges with those of three previous studies, which also observed selectivity of the outcome-to-intent shift to judgements of accidents, but not of unsuccessful attempts (Costanzo et al., 1973; Nobes et al., 2009; Zelazo et al., 1996).

Turning to children's judgments of the second story presented, however, this relationship is fully reversed (Fig. 3, bottom panel). For these judgments there is no correlation between age and the condemnation of accidental harms $r = -.10$, $p = .28$, but a strong positive correlation between age and the condemnation of attempted harms $r = .43$, $p < .001$. This was driven by low rates of condemna-

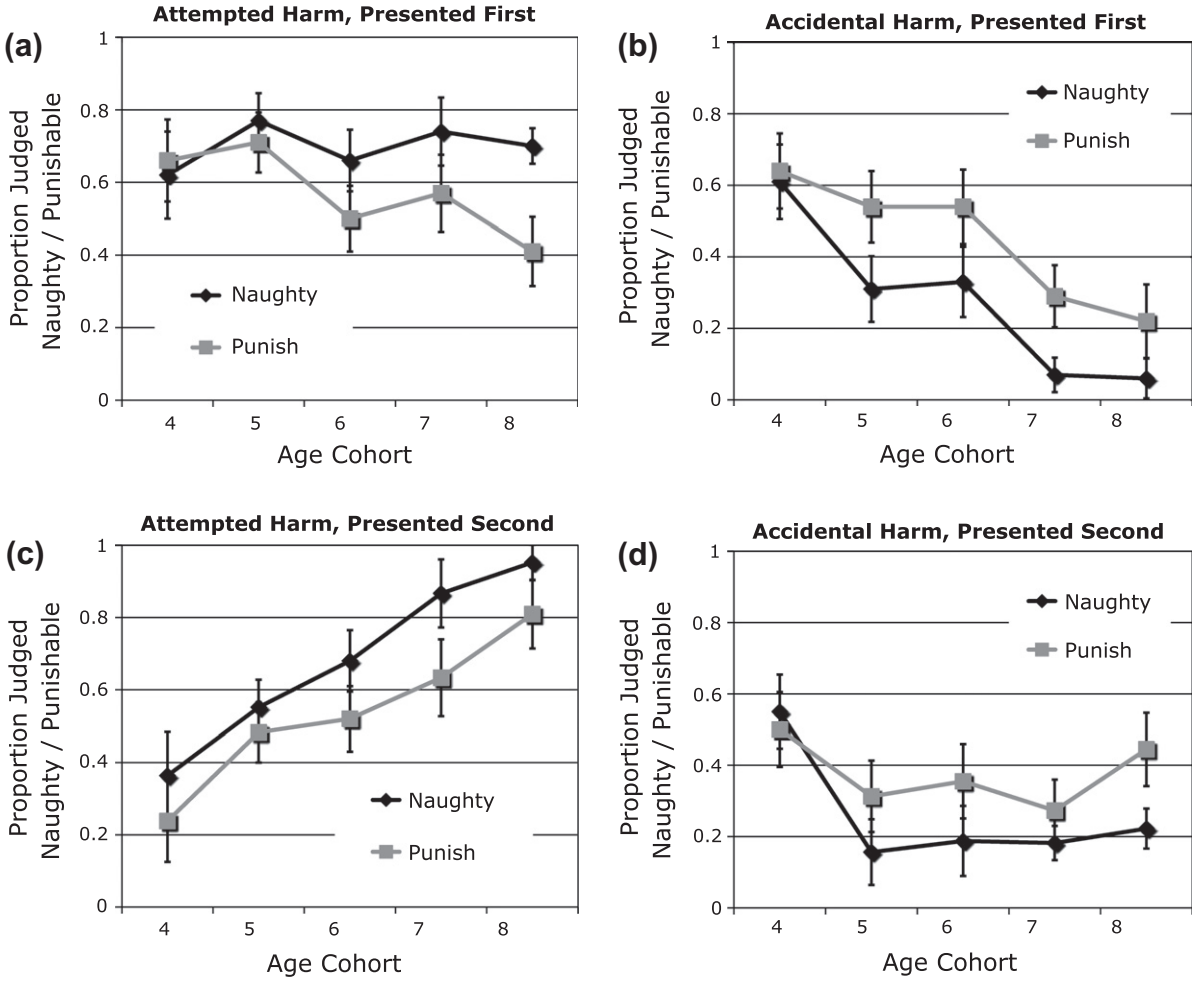

Fig. 3. Proportion of children who judged the character naughty or punishable across development, including data from responses to the first story only (top panel) and from the second story only (bottom panel).

tion of attempted harm among the youngest age groups tested.

We did not predict *a priori* that the outcome-to-intent shift would be exclusive to judgments of accidental harms for judgments of the first story, or that it would be exclusive to judgments of attempted harm for the second story. We offer a specific account for interaction by order in the general discussion. Because children's judgments are clearly susceptible to order effects, however, we will consider analyses that are restricted to first-story responses only as our primary source of data. These first-story responses can be interpreted more straightforwardly without considering the biasing role of prior information; moreover, as noted above, the patterns we observe for first-story responses are a better match to prior research (Costanzo et al., 1973; Nobes et al., 2009; Zelazo, Jacques, Burack, & Frye, 2002).

When we test the constraint hypothesis—the core prediction of the two-process theory that intent-based moral judgment emerges first for wrongness judgment and then constrains punishment judgment—we will test it exclusively for judgments of accidental harm. This is because, restricting ourselves to the more reliable and interpretable data from first-story responses, we observe evidence for the outcome-to-intent shift exclusively for cases of accidental harm. The constraint hypothesis concerns the *emergence* of intent-based moral judgment; thus, it is only sensible to test the constraint hypothesis for the subset of moral judgments where we can directly observe this emergence.

3.4. Punishment vs. naughtiness in children's judgments

The above analyses establish that the 4- and 5-year-old children follow these scenarios, can explicitly judge whether the harmful outcome occurred or not or was intended or not, and understand the questions about naughtiness and punishability. They also establish that these scenarios elicit the oft-reported shift toward greater reliance on characters' intentions with age in children's moral judgments. We now turn to our exploration of the hypothesis that the outcome-to-intent shift reflects the emergence, over these years, of the adult two-process system of moral judgments. If so, these same years should witness the adult pattern of differentiation of wrongness and punishability judgments with respect to their relative dependence on information about intent.

Adult judgments of punishment show greater sensitivity to causal responsibility for a harmful outcome than do adult judgments of moral wrongness. This pattern would be seen in our data if children judged attempted harms more naughty than punishable (because there is bad intent but no harm) and accidental harms more punishable than naughty (because there is harm, but benign intent). Fig. 3 shows this pattern to hold: the naughtiness judgments are higher than the punishability judgments for the attempted harms whereas the reverse holds for accidental harms. To establish this effect statistically, we begin by collapsing across all ages tested (4–8). In order to eliminate possible order effects we focus our analysis on each participant's first response to the first story only, thus precluding interference between stories (accidental harm vs. attempted harm) and between dependent measures (punishment vs. naughtiness). For accidental harms significantly more children judged the agent to be punishable (53%) than naughty (34%; Fisher's Exact Test $p < .05$, $N = 120$), while for attempted harms significantly more children judged the agent to be naughty (77%) than punishable (56%; Fisher Exact Test $p < .05$, $N = 133$: see Fig. 4a). Thus, like adults, children showed greater sensitivity to outcomes than to intent for punishment judgments, compared to naughtiness judgments, both in the case of accidents (bad outcome, benign intent) and attempted, but failed, harms (benign outcome, bad intent). While children tended to assign punishment at roughly the same rate to cases of accidental harm and attempted harm, our preliminary method check establishes that this pattern of response is not due to insensitivity to intent and outcome information, but instead to their roughly equal weighting.

An additional analysis combining data across all four trials (both naughtiness and punishment judgments, for both the first and second story presented) confirms this pattern of judgment (see Fig. 4b). For accidental harms, significantly more children judged the agent to be punishable (44%) than naughty (31%; Sign test: $p < .001$), while for attempted harms, significantly more children judged the agent to be naughty (75%) than punishable (59%; Sign test: $p < .001$). This is a striking pattern; one might have expected children to bring their judgments of punishment and naughtiness into perfect alignment, since each child was asked for each judgment, but even across all four questions the same pattern of judgments was observed as when only the first question on the first scenario is analyzed. In sum, children, like adults, weight outcomes more heavily in judgments of punishability than in judgments of wrongness.

Next we explore the emergence of this pattern across ages 4–8. A regression examined the effects of age and story type (attempted harm versus accidental harm) on the *difference scores* between judgments of punishment and naughtiness. This analysis asks whether the discrepancy between wrongness and punishment judgments changed between ages 4 and 8. Again we begin with first-trial responses. There was a significant effect for story type (accident versus attempt) $p < .001$, $\beta = -.31$, observable on Figs. 3 and 4. That is, this difference score was positive for accidental harm stories (+13%) and negative for attempted harm stories (−16%) indicating greater reliance on outcome information for punishment judgments. There was no main effect of age ($p = .45$, $\beta = .05$). Critically for the present analysis, there was a significant interaction between story type and age ($p = .02$, $\beta = .14$). In other words, with age, children's difference scores became more positive for accidental harm stories and more negative for attempted harm stories—in each case an effect consistent with increasing outcome bias for punishment judgments relative to naughtiness judgments. (This effect remained significant in separate analyses that included responses to the second story as well). We can calculate the full size of the discrepancy at each age by subtracting the percent judged naughty minus punishable for attempted harms, and the percent judged punishable minus naughty for acci-

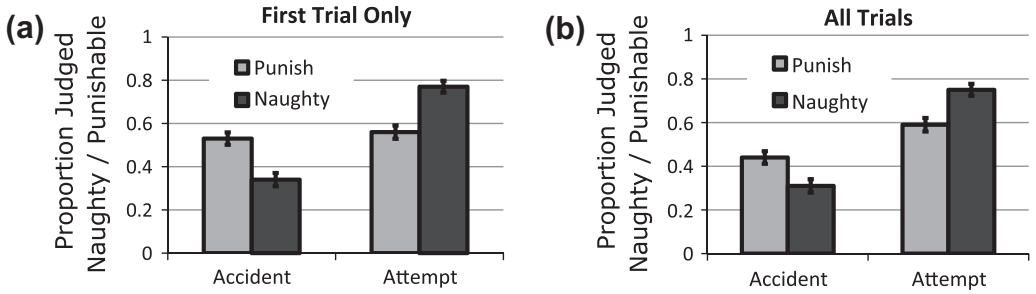

Fig. 4. Proportion of children who judged the character naughty or punishable collapsing ages 4 through 8, analyzing (a) each child's first response to the first trial only and (b) all responses to all trials.xl

dental harms, and then summing these differences. We find no discrepancy at 4 years, but statistically significant discrepancies (Fisher Exact Test) of 26% at 5, 36% at 6, 42% at 7 and 46% at 8 (Fig. 3).

Thus, as predicted by the hypothesis that the outcome-to-intent shift reflects the emergence of the adult two-process architecture of moral judgment, during the period that children exhibit increasing sensitivity to intent in their moral judgments they also begin to take outcome into account relatively more for judgments of punishment, while taking intent into account relatively more for judgments of wrongness (here, naughtiness).

The evidence from 4-year-olds must be treated cautiously; we have already noted the elevated rates of comprehension check failures among 4 year olds, indicating that our task is especially difficult for this youngest age group. Could the apparent lack of differentiation by 4 year olds be attributable to a consistency bias—that is, having answered "yes" or "no" to the first moral probe, were they more likely to offer the same answer to the second? Apparently not: the overall rate of consistency between punishment and wrongness judgments did not differ much across age groups: 84% (4 yrs), 79% (5 yrs), 74% (6 yrs), 80% (7 yrs) and 78% (8 yrs). Also, we excluded a much larger proportion of 4-year-olds due to failure of comprehension checks; could these exclusions have affected our results? Again, apparently not: Comparing these populations on their judgments of naughtiness and deserved punishment, for both accidents and attempts, all p values are above 0.35. Moreover, the proportions of 4-year-old children who judge the agent naughty or punishable, for both accidents and attempts, change by fewer than 4% points when no children are excluded from analysis.

Thus, our experiment provides some evidence that 4 year olds genuinely fail to accord greater weight to caused harm when making judgments of deserved punishment compared with judgments of moral wrongness, a pattern that emerges over the next two years and remains through adulthood.

3.5. Testing the developmental constraint hypothesis: mediation analysis

The above analyses suggest that the outcome-to-intent shift indeed reflects the emergence of the adult two process architecture of moral judgment. These analyses leave open whether its emergence reflects developments outside of the moral domain exclusively (e.g., developments in theory of mind that guarantee the relevant input to moral judgment, or developments in executive function needed to manage the occasional conflict between the outcome of each process). Alternatively, the shift may also reflect conceptual reorganization within the system of moral reasoning itself; the emergence of an intent-based process for judging moral wrongs. If the locus of the developmental change is outside of the moral domain, it should equally affect naughtiness and punishment judgments, for both require theory of mind and cognitive control to the same extent (the parallel hypothesis). In contrast, if the locus of the developmental change is first in the emergence of the intent-based moral wrongness system, naughtiness judgments should mediate the relationship between age and punishment judgments for accidental harm scenarios (the constraint hypothesis).

According to the constraint hypothesis, the correlation between age and intent-based punishment should be substantially reduced when controlling for intent-sensitivity in naughtiness judgments. However, the constraint hypothesis does not necessitate the opposite relationship: the correlation between age and intent-based naughtiness need not be substantially reduced after controlling for intent-sensitivity in punishment judgments. According to the parallel hypothesis, no asymmetry is predicted between these mediation models. Rather, the parallel hypothesis predicts that punishment judgments should explain just as much of the naughtiness judgment/age relationship as naughtiness judgments should explain of the punishment judgment/age relationship. We therefore tested both meditation models using logistic regression. We begin by analyzing responses to all trials, and then turn to a selective analysis of responses to the first story only.

Across all trials, age significantly predicted decreased judgments of the punishability of accidental harms before controlling for naughtiness judgments of accidents ($z = 2.29$, $p < .05$), but no significant relationship remained after controlling for naughtiness judgments ($z = 0.09$, $p = .93$). Testing the opposite mediation model, age significantly predicted decreased naughtiness judgments of accidents before controlling for punishability of accidents ($z = 4.45$, $p < .001$), and this significant relationship remained after controlling for punishability of accidents ($z = 3.65$, $p < .001$). We confirmed these effects by the Sobel test for mediation. Naughtiness judgments mediated 98% of the punishment judgment/age relationship ($z = 3.76$,

$p < .001$). By contrast, punishment judgments mediated only 27% of the naughtiness judgment/age relationship ($z = 2.17$, $p < .05$).

We then preformed an identical series of analyses on subjects' responses to the first story only. Age significantly predicted decreased punishability of accidents before controlling for naughtiness judgments of accidents ($z = 2.94$, $p < .005$), but not after controlling for naughtiness judgments ($z = 1.26$, $p = .21$). Testing the opposite model, age significantly predicted decreased naughtiness before controlling for punishment of accidents ($z = 4.26$, $p < .001$), and also after controlling for punishment of accidents ($z = 3.42$, $p < .005$). Again, these effects were confirmed by the Sobel test for mediation. Naughtiness judgments mediated 75% of the punishment judgment/age relationship ($z = 3.08$, $p < .005$). By contrast, punishment judgments mediated only 28% of the naughtiness judgment/age relationship ($z = 2.45$, $p < .05$).

3.6. Testing the processing constraint hypothesis: order effects

The mediation analysis presented in the previous section demonstrates that developmental changes in intent-based naughtiness judgments constrain developmental changes in judgments of whether accidental harms deserve punishment. In this section we seek complementary evidence at a processing level. Specifically, we assess order effects on the exculpation of accidents. Roughly half of our participants were first asked whether an accidental harmdoer was naughty and next asked whether he or she should be punished; the remaining half answered these questions in the opposite order. According to the constraint hypothesis, intent-based naughtiness judgments constrain whether children deem accidental harms punishable—an effect which might be particularly strong when a punishment judgment is made directly after a naughtiness judgment. Thus, participants who make naughtiness judgments first should be relatively less likely to deem accidental harms punishable, compared with those who make punishment judgments first. However, the reverse effect is not predicted by the constraint hypothesis: Participants who make punishment judgments first should be no more likely to judge accidental harmdoers naughty.

As predicted, a smaller proportion of children judged the accidental harmdoer punishable when having first judged naughtiness (30%) than when having first judged punishment (52%; Fig. 5a). In contrast, similar proportions of children judged the accidental harmoder naughty when making this judgment first (27%) compared with after making a punishment judgment (24%). We modeled these effects using logistic regression, predicting negative moral assessment of the accidental harmdoer by the judgment type (punishment versus naughtiness), order of judgment (punishment first versus naughtiness first), and their interaction. There was a significant main effect of judgment type ($z = -3.51$, $p < .001$) and a marginally significant effect of order ($z = 1.91$, $p = .056$), but critically these were qualified by a significant interaction ($z = -2.66$, $p < .008$), indicating an order effect exclusively on punishment judgments, but not naughtiness judgments. This is predicted by the constraint hypothesis.

According to the constraint hypothesis this order effect should be particularly pronounced for children aged

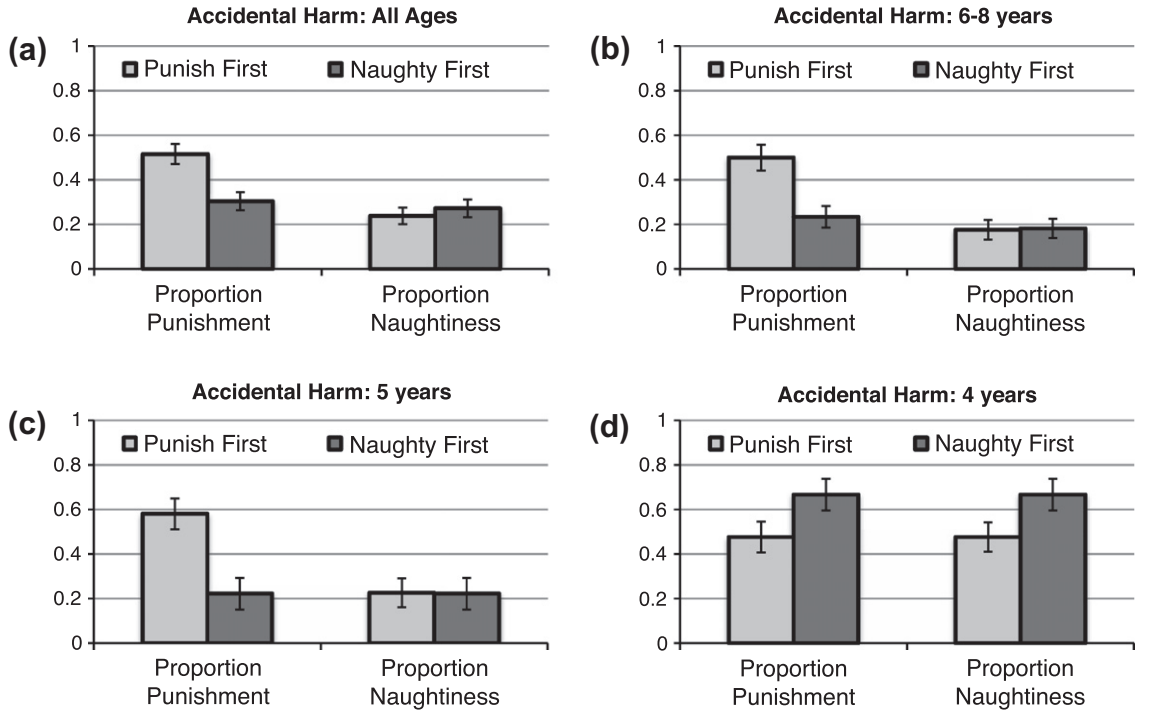

Fig. 5. Proportion of children who judged the character naughty or punishable as a function of the order in which the judgments were made (punishment first versus naughtiness first). Panel (a) shows the effect across all ages, (b) shows children aged 6–8, (c) shows 5 year olds and (c) shows 4 year olds.

5 years and older, because it is among these older children that the intent-based naughtiness judgments tend to reflect the operation of the mental state-based process of moral judgment. For children aged 4, however, many instances of "intent-based" naughtiness judgments may not reflect the developmental attainment of the mental state-based process of moral judgment. Following the design of the logistic regression reported above, we tested for an interaction between judgment type (punishment versus naughtiness) and order of judgment (punishment first versus naughtiness first) separately for different age groups. Consistent with our prediction, we found a strong and statistically significant interaction effect ($p = .026$) effect in children aged 6–8 years (Fig. 5b), an equally strong and marginally significant effect ($p = .07$) among children aged 5 years (Fig. 5c), but no evidence at all for such an effect in children aged 4 years ($p = .915$, Fig. 5d). We then tested whether these findings hold for the first story presented only; again we found a significant order $\times$ judgment type interaction across all age groups ($p = .019$), including among children aged 5 years ($p = .011$), but no evidence for this pattern of results in children aged 4 years ($p = .886$).

4. General discussion

Consistent with many prior studies we found that young children's moral judgments exhibit increasing reliance on mental states such as intent as they age. Among judgments of the first scenario presented (presumably the best measure of spontaneous judgment processes), however, there is substantial decrease in the condemnation of accidental harms, but no developmental change in the judgment of attempted harms (see also Costanzo et al., 1973; Nobes et al., 2009; Zelazo et al., 1996).

During the period when children show increasing reliance on mental state information in their moral judgments they also show increasing dissociation in the criteria for assessing naughtiness versus deserved punishment. Specifically, older children come to exhibit relatively greater reliance on mental-state information in their judgments of naughtiness of acts than in their judgments of punishability. This finding motivates the "constraint hypothesis": that intent-based moral judgment emerges in the form of a new concept of moral wrongness that subsequently constrains judgments of deserved punishment.

Two additional findings support the constraint hypothesis. For accidental harms, the effect of age on intent-based punishment judgment is mediated by intent-based naughtiness judgments, but not vice versa. And, making an intent-based naughtiness judgment strongly biases subsequent punishment judgments to be intent-based also, but not vice versa.

4.1. The two-process model of moral judgment

These findings both support and extend the two-process model of moral judgment. First, they extend a signature of the two-process model to the age of 5, with increasing strength from ages 6–8. According to that model, the causal process of moral judgment is triggered by the occurrence of a harmful outcome and assigns moral blame to the individual who is causally responsible. The mental state process of moral judgment instead assigns moral blame based on the mental states that give rise to action, especially intent to cause harm. Previous evidence suggests that among adults both processes jointly determine judgments of deserved punishment, whereas judgments of moral wrongness are almost exclusively determined by the mental state process. Our study shows developmental continuity in this pattern beginning at age 5, but no evidence for it in 4-year-olds. These data thus suggest that the outcome-to-intent shift in moral reasoning reflects, at least in part, the attainment of the adult two-process architecture underlying moral judgments.

The two-process model predicts strong conflict between processes for cases of accidental harm (because the causal process is triggered by the occurrence of harmful outcome), but not for cases of attempted harm (because the causal process is not triggered in the absence of a harmful outcome). If the outcome-to-intent shift reflects the attainment of the adult-two-process architecture, more developmental change should be observed for the accidental harm scenarios. Consistent with this prediction, we found that judgments of accidental harm underwent substantial developmental change from ages 4–8, presumably driven by the influence of an emerging mental state process over the extant causal process. Judgments of attempted harm, in contrast, changed much less with age, and not at all for the first story presented. Rather, on the first scenario, about 70% of children assigned punishment and naughtiness to attempted harms across all ages tested. This high and unchanging rate of condemnation of attempted harms—despite the absence of a harmful outcome—is consistent with the hypothesis that the causal process is silent when no harmful outcome occurs. (Of course it is a challenge to explain why very young children consider attempted harms morally bad, and we take up this challenge further below).

The finding that 5- to 8-year-olds' judgments of deserved punishment of accidental harm are constrained by their mental state-based judgments of naughtiness further supports the suggestion that the two-process system of moral judgment is coming on line between ages 4 and 6. Although this evidence alone is not sufficient to decide how adults judge the punishment deserved for accidental harms, it suggests that adult judgments of punishment for accidents may be similarly constrained by a mental state-based concept of moral wrongness.

Our evidence does not, however, suggest a rapid change occuring between the ages of 4- and 5-years-old for all children, but rather a more gradual change spanning several years and likely extending earlier than age 4 at least in some children. Like previous studies (Armsby, 1971; Farnill, 1974; Yuill & Perner, 1988), we find evidence that even 4-year-old children are sensitive to some mental-state features even for cases of accidental harm. For instance, while they condemn accidents more than do older children, nevertheless 4-year-olds condemn intentional harms even more than accidents. More strikingly, they robustly condemn attempted harms, consistent with evi-

dence from preverbal infants (Hamlin et al., in press). So, we are presented with a challenge: How can we reconcile apparent evidence for intent-based judgment from an early age with additional evidence for a reorganization of moral concepts during the preschool years that centers on the role and scope of intent-based judgments? We take up this challenge in two steps, first focusing on the evidence for conceptual reorganization, and then reconciling this with an account of early-emerging sensitivity to intent for attempted harms.

4.2. Conceptual attainments in the moral domain

The conceptual reorganization implied by our findings involves the emergence of a mental-state based conception of moral wrongness sufficient to exculpate accidental harms, the differentiation of the concept of wrongness from punishability, and ultimately the constraint of punitive judgment by wrongness judgment.

Such conceptual reorganization within the moral domain has not been emphasized in recent discussions of the outcome-to-intent shift, which attribute it to the attainment of capacities outside the moral domain such as theory of mind (Chandler et al., 2001; Killen et al., 2011) and executive function (Zelazo et al., 1996). The development of these capacities presumably do contribute to the outcome-to-intent shift in moral judgment, just as past theories and evidence indicate. Intent-based moral judgment surely requires a capacity to represent others' mental states and to connect those mental states to action, and it likely also requires executive function to negotiate the diverse outputs of two distinct processes of moral judgment.

Still, these hypotheses alone cannot explain key findings of the present study. First, neither hypothesis explains the increasing differentiation between punishment and naughtiness judgments, with respect to the relative importance of outcome versus intent, between ages 4 and 8. Second, neither hypothesis explains why intent-based moral judgments of accidental harm emerge first in judgments of moral wrongness and then subsequently constrains judgments of deserved punishment. The requirement of representing another person's mental state is equally demanding for judgments both of deserved punishment and naughtiness, as is the requirement of suppressing the causal process in favor of the mental state process of moral judgment. Whatever cognitive attainments outside the moral domain may be necessary for the outcome-to-intent shift, this shift appears also to involve a reorganization of concepts within the moral domain.

Our interpretation of the changes share some core features with Piaget's and Kohlberg's stage theories, but extend and even contradict those theories in other respects. Supporting Piaget and Kohlberg, we have presented new evidence for a reorganization of moral concepts that includes both a shift from outcome- to intent-based judgment, and from a punishment- to wrongness-based conception of moral transgression. But for Piaget and Kohlberg, those two conceptual shifts were explained by a common third cause: a domain-general cognitive shift from concrete and egocentric thought to abstract thought. We have suggested instead that these "two" conceptual shifts are in fact two faces of the same phenomenon: the emergence of a new process of moral judgment that locates the source of moral wrongness in the nature of the mental states that give rise to action.

Additionally, Piaget and Kohlberg emphasized the qualitative differences between successive stages, with the interrelated concepts of later stages fully replacing those of earlier ones. In contrast, we have documented striking continuities in the two-process model of judgment, with adult-like structure beginning to emerge by 5 years old. Furthermore, we have emphasized the continuous presence of one process—the outcome based process—from early preschool years through adulthood. We agree that the second, mental-state based process arises from an important reorganization of the conceptual underpinnings of moral reasoning. But when this new process of judgment is attained it does not replace the outcome process as a successive stage; rather, both processes coexist into adulthood, and their competitive interaction explains the dilemma of moral luck.

Finally, Piaget and Kohlberg were concerned with children's explicit, verbalizable, moral theories. This focus is particularly clear in Kohlberg's work, which diagnoses children's explicit concepts from their justifications for their moral judgments rather than from patterns of judgments themselves. More recent research challenges this approach, indicating that adult moral judgment is largely the product of automatic and intuitive psychological mechanisms (Cushman et al., 2006; Greene, 2008; Haidt, 2001). Our study does not speak directly to whether the causal and mental state processes of moral judgment are best characterized as explicit theories or automatic processes (or both). Nevertheless, (1) the developmental persistence of the causal process of moral judgment into adulthood, (2) the fact that it is constrained, and not replaced, by a mental state-based concept of moral wrongness, and (3) evidence for counterinterive phenomena such as blame-blocking, all suggest that the causal process is at least partly the product of automatic processes of moral judgment in adults and children alike. (See also a recent report by Buon and colleagues (2013) in this journal).

Our experimental method relied exclusively on the analysis of children's judgments, and we did not solicit or analyze their verbal reports of underlying reasoning processes. This conservative approach was motivated by the concern that children's verbal reports, like adults', might not accurately reflect structure of automatic processes or implicit concepts in the moral domain (Cushman et al., 2006; Haidt, 2001; Nisbett & Wilson, 1977). A valuable area for future research, however, is to directly query children's reasoning processes in order to establish the elements of moral judgment that are, and are not, a product of controlled reasoning.

4.3. Sensitivity to intent in early childhood

Two salient aspects of our data were not predicted *a priori* and are not straightforwardly explained by the model of conceptual attainment we offered above. First, we found that even most 4 year old children condemned attempted harms, when analyzing responses to the first story pre-

sented. Second, the locus of developmental change reversed when analyzing the second story presented, for which we found development change for judgments of attempted harm, but not for judgments of accidental harm. Because our experiment was not designed to test for these effects our accounts of these phenomena are necessarily tentative. Nevertheless, we think they offer important insight into the psychological mechanisms that support moral judgment in children and adults.

We begin by focusing on moral judgments of the first story presented, which presumably offer an unbiased view of the cognitive mechanisms that children spontaneously deploy when making moral judgments. Our model of conceptual attainment accounts for developmental change in the judgment of accidental harms; the challenge is to explain developmental continuity in the judgment of attempted harms, and in particular the high level of condemnation of harmful attempts even at a young age. Consistent with this finding, there is other evidence that preschoolers (and even infants) take intent into account in their evaluations of others' actions (Armsby, 1971; Farnill, 1974; Hamlin et al., in press; Nobes et al., 2009; Yuill & Perner, 1988). What psychological process causes a four year-old child to condemn attempted harmdoing, but is *not* sufficient to consistently exculpate accidental harmdoing (as would a full mental-state based process of moral judgment)?

We suggest that children have an early developing automatic negative reaction to "bad acts", along the same lines as their capacity for their negative reaction to "bad outcomes" that gets harnessed during the preschool years to concepts of naughtiness, punishability, and wrongness (concepts that are not initially differentiated). At this point they do not have a broad concept of moral wrongness according to which intentional action occupies a central role. This suggestion resonates with Kohlberg's characterization of early-stage moral reasoning as focused on "external, quasi-physical happenings [and] bad acts". We and others have emphasized the young child's reliance on outcomes (i.e., external/physical happenings) as a basis for moral judgment at these ages, but less attention has been paid to the identification of bad acts.

One basis for the identification of bad acts, closer to Kohlberg's original proposal, might involve discrete categorizations of particular action (e.g., "pushing") based on both their motor properties (pushing = lunging at person with hands out) and goal properties (pushing = action with goal of knocking to the ground). Consistent with this possibility, even infants represent actions in terms of their yet-unrealized goals (Meltzoff, 1995; Woodward & Sommerville, 2000). In other words, what triggers condemnation may be a representation of the form "pushing person action" or "breaking mirror action", which are proscribed as particular actions, rather than the abstract notion "action performed with intent to harm", which is proscribed on the general basis of the intent to harm.

A conceptually richer basis for the identification of bad acts is, simply, malicious intent. In other words, what triggers condemnation may not be that the action is a proscribed "pushing" action, but rather that the action was motivated by an intent to cause harm. Consistent with this richer basis, recent evidence shows that preverbal infants appear to negatively evaluate agents who attempt, but fail, to harm another by preventing them from achieving a goal (Hamlin, under review). There are merits to both the leaner and richer versions, and they are not mutually exclusive. If the richer version is correct, however, it constitutes a process of moral judgment (or at least agent evaluation) that is early-emerging and exhibits sensitivity to mental states, yet is distinct from the mental-state-based concept of moral wrongness that later exculpates accidents and constrains judgments of punishment. Our aim is to sharpen that distinction.

Critically, the process of identifying bad acts is posited to operate as a kind of "feature detector": when a bad act is detected, negative evaluation is triggered (Fig. 6, left panel). In this sense, the process is analogous to that proposed for outcome-based judgment, which acts as a feature detector for bad outcomes. A bad act detector would be silent when a bad act is absent (as in an accident), just as the outcome-based process is silent when a harmful outcome is absent (as in a failed attempt). Thus, consistent exculpation of accidental harms awaits the emergence of a full mental-state-based concept of moral wrongness—one that regards the absence of harmful intent as meaningful because intent is central to the concept of wrongful action (Fig. 6, right panel).

There is some evidence for automatic moral condemnation based on "bad acts" in adult moral judgment. Most people are more willing to endorse stopping an out-of-control trolley car by dropping a person in front of it with a switch, as compared to pushing a person in front of it with their hands (Cushman et al., 2006; Greene et al., 2009). Clearly the outcomes are identical in either case, but the situations differ in that one of them requires a canonically bad act (a "pushing person" act). Similarly, people show physiological signs of aversion to performing pretend harmful actions; for instance, hitting a plastic baby doll against a table (Cushman, Gray, Gaffey, & Mendes, 2012). In the absence of either a harmful outcome or any intent to harm, their aversive response appears to be tied to the performance of a harmful action itself. Notably, damage to the ventromedial prefrontal cortex is associated with decreased condemnation of attempted harm (Young, Bechara et al., 2010), and this same brain region is implicated in the affective prohibition of canonical harmful actions like pushing a person in front of a train (Ciaramelli, Muccioli, Ladavas, & di Pellegrino, 2007; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Koenigs et al., 2007).

Thus, just as moral condemnation based on causal responsibility for a harmful outcome persists into adulthood, there is also evidence moral condemnation based on the identity of the action persists into adulthood. And, just as the developmental persistence of the "bad outcome" detector leads to conflict with a mental-state based concept of wrongful action—and thus the philosophical dilemma of moral luck—the developmental persistence of the "bad acts" detector may lead to conflict with a mental-state based concept of moral wrongness in philosophical dilemmas like the trolley problem. The intent to save as many lives as possible is unimpeachable, yet the action of throwing a man in front of a train is strongly identified as bad.

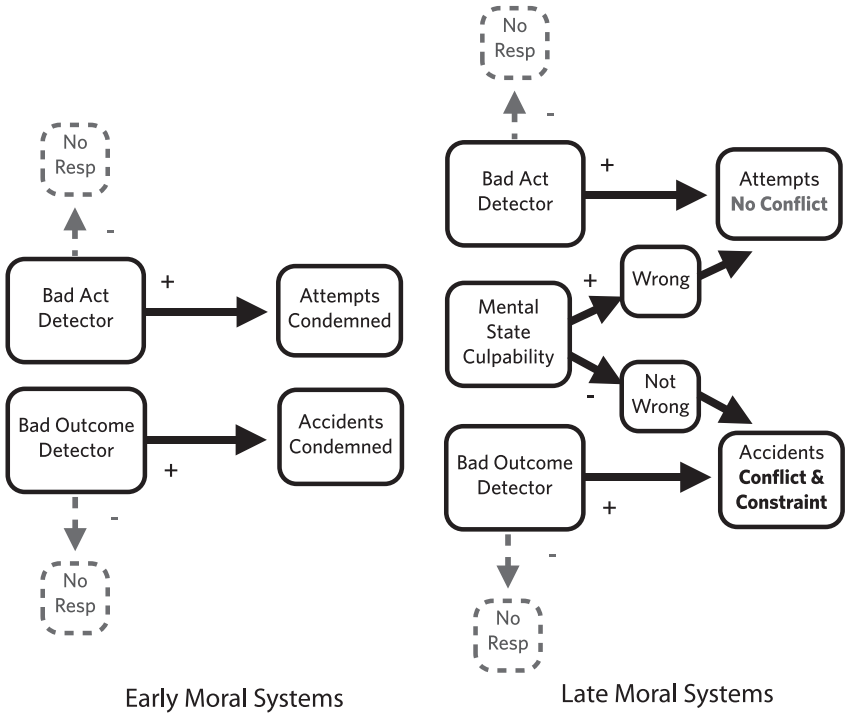

Fig. 6. A model of the early and late moral systems incorporating the bad acts hypothesis.

Two studies of young children's moral judgments provide evidence directly relevant to the bad acts hypothesis, but with inconsistent results. Contra the bad acts hypothesis, Zelazo et al. (1996) show that young children mostly consider it acceptable to hit an animal that likes to be hit. However, supporting the bad acts hypothesis, Weisberg and Leslie (2012) show that young children mostly consider it wrong to hit a person who cannot feel hurt and never cries. Resolving this apparent inconsistency is an important matter requiring further investigation.

In summary, we suggest that four year-olds tend to condemn accidental harms because of the presence of a bad outcome, while they tend to condemn attempted harms because of the presence of a bad act. Even for infants, mental states are essential to the identification and categorization of actions, and hence play a role in children's reasoning about bad acts. As children attain a new concept of moral wrongness predicated on culpable mental states like intent to harm this leads to a change in their judgment of accidental harms (which are not accompanied by a culpable mental state), but no change in their judgment of attempted harms (which are). This account fits the pattern of evidence obtained for children's responses to the first story presented in our experiment.

How can we explain the starkly different patterns of moral judgment obtained for the *second* story presented? We assume that order effects generally arise because the process of judging the first stimulus highlights a factor that subsequently takes on additional influence in the judgment of the second stimulus. We have suggested that very young children condemn cases of accidental harm because they focus on the harmful outcome. Possibly, these children are subsequently less likely to condemn a case of attempted because they focus on the *absence* of a harmful outcome (the feature highlighted in the first judgment) rather than the presence of a bad act. Similarly, we have suggested that very young children condemn cases of attempted harm because they focus on the bad act performed. Possibly, these children are subsequently less likely to condemn a case of accidental harm because they focus on the absence of a bad act (the feature highlighted in the first judgment), rather than the presence of a bad outcome. Finally, over the ages 5–8, children are hypothesized to begin to exculpate accidental harmdoers on the basis of their lack of intent to harm, and thus may become hyper-sensitive to the presence of intent when subsequently judging the attempted harmdoer, leading to the observed increase in condemnation of the attempted harmdoer in this older age group (Fig. 5, bottom left panel). In each case we posit a similar mechanism: Judgment of the first case highlights a salient feature, and the presence or absence of that feature dominates judgments of the second case. What changes over development, of course, are the features made salient by the child's available processes of moral judgment.

4.4. Developmental change in the attribution of negligence?

A recent report by Nobes et al. (2009) suggests an alternative interpretation of our data that deserves special attention. Nobes and colleagues propose that by 3 years of age the basic mental-state process of moral judgment deployed for both accidental and attempted harms are in place, and that apparent developmental change is due to an over-application of the concept of negligence among very young children. They propose that very young children condemn accidents because they infer that bad out-

comes must be caused by careless behavior, whereas older children and adults allow that bad outcomes can occur even without carelessness. Two key pieces of evidence support Nobes and colleagues' conclusions. First, they find that children are less likely to condemn accidents when it is explicitly stated that the agent in question took great care to avoid the accident, and more likely when it is explicitly stated that the agent acted carelessly. Second, they find that many children refer to the carefulness or carelessness of an agent in the explicit justifications offered for their moral judgments.

For several reasons it is difficult to directly compare our results with those of Nobes and colleagues. They manipulate the degree of care taken by each agent, which is directly stated in their stimuli, but we do not. Their study and ours also manipulate intent (vs. outcome) quite differently. We manipulate each agent's desire (e.g. wanting to knock somebody to the ground) independent of their outcome (e.g. actually knocking somebody to the ground). In contrast, Nobes and colleagues typically manipulate an antecedent action that they label the "intention" (e.g. stealing vs. owning a bike) independent of a subsequent action (e.g. crashing the bike). Thus, while they find very strong sensitivity to their intent factor even at 3–4 years old (and use this to argue for conceptual continuity), this may be because "intent" was often instantiated as a fully completed and independently bad action, such as stealing a bike.

Concerning the evidence for negligence-based judgment, Nobes and colleagues' cases of "careless" action often involve an *internal* cause while their "careful" actions involve an *external* cause. For instance, in one story a girl drops a puppy, hurting it, either because she is only holding it with one hand (careless) or because the puppy jumps despite the girl holding on securely (careful). Preschoolers are less likely to condemn this accident in the careful case, which Nobes and colleagues attribute to an inference about the lack of negligence. An alternative interpretation is that people do not attribute causal responsibility to the girl when an external force—the puppy jumping—brings about the harmful outcome. This would be consistent with young children making judgments of accidental harms largely by assessing causal responsibility for the harm, rather than carelessness. Regarding children's explicit justifications, it is possible that they appealed to careful versus careless behavior because these stimulus manipulations were made highly salient and reinforced during comprehension probes (although outcome information was, as well). Indeed, ample evidence suggests that even adults' explicit justifications fail to capture the underlying basis of their moral judgments (Cushman et al., 2006; Haidt, 2001; Nisbett & Wilson, 1977). Still, at a minimum this evidence indicates that young children are sufficiently sensitive to carelessness as a morally-relevant dimension to appeal to it in their explicit justifications.

The evidence presented by Nobes and colleagues has considerable weight, and an important area for future research is to reconcile the divergent methodologies employed in these two studies in order to also reconcile their differing interpretations. Several factors motivate our claim that young children's moral judgments are determined largely by assessing causal responsibility, and not merely by overattributing carelessness as Nobes and colleagues argue. Our proposal provides an explanation for the evidence that moral wrongness judgments comes to constrain punishment judgments concurrent with the exculpation of accidental harm. And, our proposal links developmental changes in the moral judgment of children with a two-process architecture supported by studies of adults. Elements of the adult data, such as blame blocking (Cushman, 2008) and the condemnation of outcomes caused purely by chance (Cushman et al., 2009), are not easily accounted for by a negligence-based model.

4.5. Conclusion

Our study provides evidence for a change in the processes by which children make moral judgments over the years of 4 to 8, suggesting a reinterpretation of the classic stage theories of Piaget and Kohlberg that incorporates recent work on adult moral judgment. From a very young age children condemn actions that cause harm; this process appears to be relatively automatic and continuous through adulthood. In addition, our findings provide tentative evidence that young children may similarly condemn bad acts, like intentionally shoving a person to the ground. Around 6 years old, children acquire a new process of moral judgment according to which actions are condemned on the basis of mental states such as intent to harm and foresight of harm. This mental-state process of judging moral wrongness leads to the exculpation of accidental harms, and also constrains judgments of punishability. The joint operation of these two psychological processes—the early-emerging condemnation of harm caused, and the later-emerging condemnation of harm intended—form the basis of the adult two process architecture, giving rise to the surprising and confounding dilemma of moral luck.

The question of exactly why, and how, the adult two-process structure is attained over these years is open. Additionally, the present study and most of the adult studies that motivate it have focused on cases of harmful action; it is an open question whether the same processes govern the judgment of helpful or praiseworthy action. It is also uncertain whether each process is best characterized as an automatic or controlled process of moral judgment. To address this question, studies eliciting explicit justifications for the judgments obtained here would provide relevant data. Finally, much remains to be learned about the process of identifying "bad acts" that apparently supports the condemnation of attempted harms from a young age. The significance of the present analysis derives from the developmental evidence it provides that converges with the adult evidence for the two process theory of moral judgment, and the insight it gives into the classic outcome-to-intent shift in moral reasoning over the years of 4–8.

Acknowledgements

We thank the Museum of Science in Boston, especially Marta Biarnes and the staff of the Discovery Center, for facilitating this research. We also thank Kiley Hamlin, Marc Hauser, Gavin Nobes, David Sobel and Liane Young for valuable feedback and suggestions.

References

Armsby, R. E. (1971). A reexamination of the development of moral judgments in children. *Child Development*, 1241–1248.

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, 103*, 37–49.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Berg-Cross, L. (1975). Intentionality, degree of damage, and moral judgments. *Child Development, 46*(4), 970–974.

Buon, M., Jacob, P., Loissel, E., & Dupoux, E. (2013). A non-mentalistic cause-based heuristic in human social evaluations. *Cognition, 126*(2), 149–155.

Carey, S. (1985). Are children fundamentally different thinkers and learners from adults? In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.). *Thinking and learning skills* (Vol. 2). Hillsdale, NJ: Earlbaum.

Chandler, M. J., Sokol, B. W., & Hallett, D. (2001). Moral responsibility and the interpretive turn: Children's changing conceptions of truth and rightness. *Intentions and Intentionality: Foundations of Social Cognition*, 345–365.

Ciaramelli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive Affective Neuroscience, 2*, 84–92.

Costanzo, P., Coie, J., Grumet, J., & Farnill, D. (1973). A reexamination of the effects of intent and consequence on children's moral judgments. *Child Development, 44*(1), 154–161.

Cushman, F. A. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380.

Cushman, F. A., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a 'trembling hand' game. *PLOS One, 4*(8), e6699. doi:6610.1371/journal.pone.0006699.

Cushman, F. A., Gray, K., Gaffey, A., & Mendes, W. (2012). Simulating murder: The aversion to harmful action. *Emotion, 12*(1), 2–7.

Cushman, F. A., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience, 7*(3).

Cushman, F. A., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17*(12), 1082–1089.

Darley, J. M., & Shultz, T. R. (1990). Moral rules – their content and acquisition. *Annual Review of Psychology, 41*, 525–556.

Farnill, D. (1974). The effects of social judgment set on children's use of intent information. *Journal of Personality, 42*(2), 276–289.

Fincham, F. D., & Jaspers, J. (1979). Attribution of responsibility to the self and other in children and adults. *Journal of Personality and Social Psychology, 37*(9), 1589–1602.

Gelman, R., & Baillargeon, R. (1983). A review of some Piagetian concepts. *Handbook of Child Psychology, 3*, 167–230.

Gino, F., Moore, D., & Bazerman, M. (2008). No Harm, no foul: The outcome bias in ethical judgments.

Gino, F., Shu, L., & Bazerman, M. (2010). Nameless + harmless = blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior. *Organizational Behavior and Human Decision Processes, 111*(2), 93–101.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.). *Moral psychology* (Vol. 3). Cambridge, MA: MIT Press.

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*(3), 364–371.

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

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

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Hamlin, J. K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C., (in press). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science.*

Hamlin, K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*, 557–559.

Hebble, P. W. (1971). Development of elementary school children's judgment of intent. *Child Development, 42*(4), 583–588.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Imamoglu, E. O. (1975). Children's awareness and usage of intention cues. *Child Development, 46*(39–45).

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition, 119*(2), 197–215.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908–911.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 151–235). New York: Academic Press.

Mazzocco, P., Alicke, M., & Davis, T. (2004). On the robustness of outcome bias: No constraint by prior culpability. *Basic and Applied Social Psychology, 26*(2), 131–146.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31*(5), 838.

Mikhail, J. (2002). *Aspects of a theory of moral cognition*. Public Law and Legal Theory Research Paper Series. Georgetown University Law Center.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Science, 11*(4), 143–152.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., et al. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences, 108*(7), 2688.

Nagel, T. (1979). *Mortal questions*. Cambridge: Cambridge University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259.

Nobes, G., Panagiotaki, G., & Pawson, C. (2009). The influence of negligence, intention and outcome on children's moral judgments. *Journal of Experimental Child Psychology, 104*, 382–397.

Piaget, J. (1965/1932). *The moral judgment of the child*. New York: Free Press.

Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development, 57*(1), 177–184.

Shweder, D., & Haidt, J. (1993). The future of moral psychology: Truth, intuition, and the pluralist way. *Psychological Science, 4*, 360–365.

Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child Development*, 1333–1336.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York: Guilford Press.

Weisberg, D., & Leslie, A. M. (2012). The role of victims' emotions in preschoolers' moral judgments. *Review of Philosophy and Psychology, 3*(3), 439–455.

Williams, B. (1981). *Moral luck*. Cambridge: Cambridge University Press.

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science, 11*(1), 73.

Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M., & Damasio, A. (2010). Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron, 65*(6), 845–851.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences, 107*(15), 6753.

Young, L., Cushman, F. A., Hauser, M. D., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, 104*(20), 8235–8240.

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actors responsibility and recipients emotional reaction. *Developmental Psychology, 24*(3), 358–365.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development, 67*(5), 2478–2492.

Zelazo, P. D., Jacques, S., Burack, J. A., & Frye, D. (2002). The relation between theory of mind and rule use: Evidence from persons with autism-spectrum disorders. *Infant and Child Development, 11*, 171–195.