Cognition 167 (2017) 191–200

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Moral learning as intuitive theory revision ☆

Marjorie Rhodes [a],*, Henry Wellman [b]

[a] New York University, United States
[b] The University of Michigan, United States

ARTICLE INFO

Article history:
Received 18 January 2016
Revised 29 June 2016
Accepted 24 August 2016
Available online 6 September 2016

Keywords:
Moral learning
Intuitive theories
Conceptual development

ABSTRACT

We argue that moral learning, like much of conceptual development more generally, involves development and change in children's intuitive theories of the world. Children's intuitive theories involve coherent and abstract representations of the world, which point to domain-specific, unobservable causal-explanatory entities. From this perspective, children rely on intuitive sociological theories (in particular, an abstract expectation that group memberships constrain people's obligations), and their intuitive psychological theories (including expectations that mental states motivate individual behavior) to predict, explain, and evaluate morally-relevant action. Thus, moral learning involves development and change in each of these theories of the world across childhood, as well as developmental change in how children integrate information from these two intuitive theories. This perspective is supported by a series of research studies on young children's moral reasoning and learning, and compared to other developmental approaches, including more traditional forms of constructivism and more recent nativist perspectives.

1. Moral learning as informed by children's developing theories of agents and groups

Imagine a childhood moral transgression: A child sneaks into her classroom while she is supposed to be at recess, takes a cookie belonging to another child that she eyed during snack time, puts it in her own backpack, and leaves the room to rejoin her class. Later, when the class is asked if anyone knows where the missing cookie might be, she remains silent.

Most would agree that the child's actions were morally wrong. Broadly, several types of information feed into this judgment. At the least: (a) that the other child was harmed (left sad and hungry, with certain property rights violated), and (b) that the agent's mental states (e.g., her knowledge that the cookie belonged to someone else and her intent to take it for herself) make her culpable for these outcomes.

What sort of learning and development does such a system of moral judgment require, enable, and manifest? We view moral development, like much of conceptual development more generally, as involving the development of children's intuitive theories of the world (Gopnik & Wellman, 2012; Wellman & Gelman, 1992). On this view, conceptual structures take the form of everyday theories (Murphy & Medin, 1985), and cognitive development may be understood as a process of theory revision. Thus, via processes of constructivist learning (Gopnik & Wellman, 2012; Xu, 2007) children acquire intuitive theories of the world, revise those theories in response to new evidence, and employ those theories to learn further information. Children's intuitive theories involve coherent and abstract representations of the world, which point to domain-specific, unobservable causal-explanatory entities (e.g., gravity in the case of intuitive physics, desires in the case of intuitive psychology). Children's theories are also hierarchical—specific theories of how things work (e.g., that cookies are more desirable than carrots and the child above desires cookies) are embedded in more abstracted "framework theories" of the relevant domain (e.g., that unobservable mental states such as desires generally motivate behavior; Carey, 2009; Wellman, 1990; Wellman & Gelman, 1992). Often, children first construct a framework theory of a domain—a broad view that human behavior relies on unobservable mental states in the case of intuitive psychology. These framework theories underlie more specific theories within the domain—e.g., that desirable cookies cause specific sorts of behavior.

Development and learning can involve change in both of these levels of children's knowledge. For example, change in children's framework theories could (and does, see e.g., Wellman, 2014) include a change from the theory that desires motivate behavior to a more complex theory that the influence of desire on behavior

☆ Preparation of this article was supported by National Science Foundation grant BCS -1226942 to Rhodes and a grant from the Varieties of Understanding Project at Fordham University and the John Templeton Foundation to Wellman.

* Corresponding author at: 6 Washington Place, 301, New York, NY 10003, United States.

E-mail address: marjorie.rhodes@nyu.edu (M. Rhodes).

http://dx.doi.org/10.1016/j.cognition.2016.08.013
0010-0277/© 2016 Elsevier B.V. All rights reserved.

is moderated by beliefs and knowledge. Change in a child's more specific theory might include learning that both cookies and carrots could be desirable for different reasons (for taste or health), and that an agent's choice might reflect multiple concerns.

Children's theories serve specific cognitive functions—they enable children to predict, explain, and evaluate events in their environment (Gopnik & Wellman, 1994, 2012). For example, in the case of intuitive psychology, children can use a desire-based theory to predict that an agent will reach for the snack they desire, to then explain the agent's action (e.g., he took that one *because* he wanted it) and to evaluate whether an observed outcome was consistent or inconsistent with one's expectation. This type of evaluation allows theories to be dynamic as well—intuitive theories can change in response to observed evidence. This change often happens in a gradual and progressive manner, instead of all-at-once. For example, in the case of intuitive psychology, children move from a fairly rudimentary desire-based theory to a full-fledged representational theory of mind by passing through levels (Wellman & Liu, 2004) where they come to understand desires as moderated by simple aspects of perception, then to understand knowledge and ignorance, and then finally to understand fully the representational nature of belief (including that beliefs can be inconsistent with reality). Viewing cognitive development as a process of intuitive-theory-change encourages researchers to examine several types of development and learning. In particular, it is important to examine both how intuitive theories develop and change over extended periods of time, as well as how intuitive theories direct attention and memory to shape learning in children's day-to-day interactions. Moreover, an intuitive theories perspective encourages researchers to examine important progressions in children's conceptual development where earlier understandings within a progression set the stage for and constrain acquisition of later conceptual understandings.

Viewing cognitive development in this way contrasts with other theoretical proposals. In what follows we will distinguish our position in particular from nativist accounts, which emphasize early (evolved) understandings rather than the processes that underlie change over human development. We also distinguish the type of learning that we describe from other constructivist or social learning accounts, which describe change as motivated solely by children's responses to their own actions or from direct instruction. Finally, we distinguish our account from Social Domain Theory, which describes the cognitive domains relevant to moral judgment much differently than we propose here.

An intuitive theories perspective has been fruitfully applied to multiple conceptual domains, including intuitive physics (Kushnir & Gopnik, 2007), biology (e.g., Carey, 1985; Gelman, 2003; Inagaki & Hatano, 2002), psychology (Gopnik & Meltzoff, 1997; Wellman, 2014), and sociology (Hirschfeld, 1996; Rhodes, 2012). We view moral judgment, and thus moral learning, as resting on the interplay of intuitive psychology and sociology. Intuitive psychology shapes children's beliefs about how individuals' mental states (e.g., beliefs, desires, knowledge, traits, and so on) predict and explain behavior, whereas intuitive sociology shapes their understanding of how people relate to one another. Because moral judgments integrate information represented by both of these intuitive theories, moral learning can entail change in the relevant components of the theories that compose children's intuitive psychology or sociology, as well as in how children integrate information from these two theories. We predominantly focus on children's explicit conceptual understandings in these domains, as we will clarify in what follows.

1.1. Intuitive psychology

Return to the scenario at the beginning of this paper, but with an entirely different set of mental states. In this new account, the girl did not know the cookie belonged to another child, but simply sees it sitting on a table and thinks that it is for anyone in the class. Perhaps also, the girl doesn't realize when the teacher asks about the other child's cookie that it is the same cookie she has taken. Given these different beliefs, the agent might not be judged as morally culpable (or at least not nearly to the same extent)—her actions still caused harm, but she didn't hold the mental states necessary to make her responsible for their outcome. Or consider this scenario—the girl sees her friend with a cookie at snack time which she wants for herself, returns to the classroom to take it, but in her absence does not know that her friend in fact ate her own cookie, and what is now on the desk is a cookie that is available for anyone in the class. In this case, she had *malicious intent*, even though her mistaken beliefs and knowledge were such that her actions did not actually cause harm or infringe on anyone's property rights. Most would agree that her actions in this case were morally suspect, despite the lack of a harmful outcome.

These examples illustrate that an agent's mental states matter a great deal in everyday moral judgments. Indeed, the role of mental states in determining moral culpabilities is reflected in the legal system (e.g., in the difference between murder and manslaughter, Hart, 1968; Mikhail, 2007), and is readily recognized by adult participants in psychological studies. Adults view agents who cause harm intentionally or who intend to cause harm but fail to (because of mistaken knowledge or beliefs) as more morally culpable than those who cause harm accidentally (while trying to do good; Cushman, 2008; Knobe, 2005; Singer, Kiebel, Winston, Dolan, & Frith, 2004; Young, Cushman, Hauser, & Saxe, 2007). Further, disruption to brain regions that support reasoning about others' psychological states disrupts this pattern of judgment, leading people in this case to hold others more responsible for actions that they do not bring about on purpose (Young et al., 2007).

Clearly then, one key candidate for important developmental change and learning in moral judgments concerns children's intuitive psychologies—if there is important developmental change in children's abilities to represent and track things like knowledge, intent, and beliefs, then this will correspond to developmental change in moral judgment. Indeed, substantial developmental change occurs in the extent to which children incorporate concepts like *beliefs* and *knowledge* into their intuitive psychological theories across childhood (Wellman, Cross, & Watson, 2001; Wellman & Liu, 2004).

Detailed empirical findings provide evidence for the hypothesized process by which this type of theory-change occurs. For intuitive psychology, Rhodes and Wellman (2013) combined developmental scaling and experimental, microgenetic methods to examine the processes underlying the acquisition of a representational theory of mind within a progression of conceptual development. In employing scaling methods, they first assessed children's initial psychological theories for the extent to which children understood that (a) people have unique desires, (b) people have unique beliefs, (c) people only know what they have access to, and (d) people can believe things that are false (inconsistent with reality). Such conceptions exhibit a developmental progression in intuitive psychological understanding, proceeding from (a) to (d) as revealed in children's task performance on a well-validated developmental scale (e.g., Wellman, Fang, & Peterson, 2011; Wellman & Liu, 2004). Note that in this scale, and within the wider literature on children's intuitive psychology, an understanding of false beliefs constitutes a milestone achievement. In particular, the ability to pass an explicit false belief tasks (exemplified on the left side of Fig. 1) has often been taken to indicate the development of a full-fledged representational theory of mind (see the meta-analysis by Wellman et al., 2001).

In Rhodes and Wellman (2013), only children who initially failed explicit measures of false belief understanding, and thus

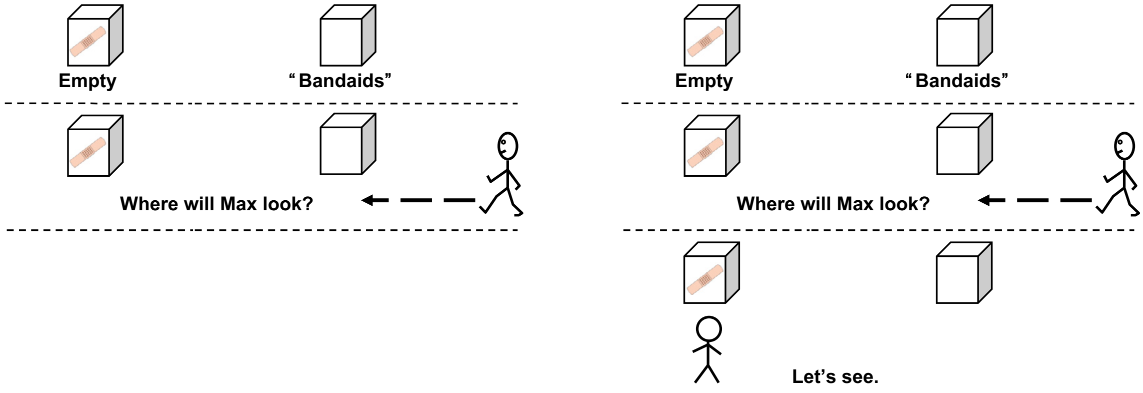

Fig. 1. A sample explicit false-belief task of the sort used in Rhodes and Wellman (2013). On the left children predict where the character will look. On the right children predict and then receive implicit feedback about their predictions, by seeing where the character actually looks. In Rhodes & Wellman, children were also often required to explain this action (e.g. "Why is Max going there?"), but were given no other feedback, including no feedback on the explanations they generated.

lacked fully representational theories of mind, continued on in the study. These children varied, however, in how close they were to achieving this understanding. In particular, some children had already obtained the conceptual insight that usually directly precedes false belief understanding (e.g., an understanding of knowledge/ignorance) and some were farther back in this conceptual progression.

Children then participated in a six-week microgenetic, experimental study, in which some children were provided with extensive evidence of the representational nature of beliefs. Microgenetic studies track participants over many closely spaced longitudinal sessions to achieve a rich picture of development and change. In the Rhodes and Wellman study, children in the focal conditions completed two false beliefs tasks per session in two sessions a week over the course of six weeks, in which they saw people acting based on false beliefs (instead of based on reality or desires) and were asked to explain why people acted the way they did. Thus, children saw evidence that was inconsistent with a simple, desire-based theory (as on the right side of Fig. 1), and were prompted to puzzle over this new evidence. The key question was whether children's theories would change based on exposure to extensive evidence indicative of false beliefs (e.g., the agent going where a desired object is *not*, due to mistaken beliefs about where it is), and in particular, if any such learning would depend on children's previous level of conceptual understanding. Indeed, this work showed that only those children who were conceptually close to developing an understanding of false belief (those who already understood knowledge vs. ignorance) underwent conceptual change following exposure to relevant evidence. These children developed an understanding of false belief by the end of the experimental period, whereas those who were farther from this understanding did not (nor did children in control conditions, regardless of how close they were to this understanding at the start). These data demonstrate how children's initial theories enable and constrain the extent to which they learn from new evidence to obtain new conceptual insights. Being closer to the focal understanding enabled children's conceptual change; being farther away constrained it. Moreover, although those children further from false-belief understanding at start failed to achieve false-belief understanding, they nonetheless progressed—they were closer to that understanding along the sequence of the ToM Scale progressions at post-test than they had been at pre-test. Thus, as outlined by an intuitive theory account, the findings show how prior knowledge influences whether exposure to new evidence results in conceptual learning. Moreover, the findings showed that such learning proceeds in orderly conceptual progressions.

Although infants show some implicit understanding of representational mental states in infancy (Onishi & Baillargeon, 2005; Scott & Baillargeon, 2009; Scott, Baillargeon, Song, & Leslie, 2010; Song & Baillargeon, 2008; Song, Onishi, Baillargeon, & Fisher, 2008; Surian, Caldi, & Sperber, 2007), as we will discuss later, the development of an explicit representational theory of mind in preschool remains an important conceptual achievement. Development of an explicit theory of mind in the preschool years qualitatively changes how children interact with their environment, influencing children's social competence (Astington, 2003; Peterson & Siegal, 2002; Peterson, Slaughter, & Paynter, 2007; Slaughter, Imuta, Peterson, & Henry, 2015), peer interactions (Dunn, Cutting, & Demetriou, 2000), and capacities to engage in behaviors like pretense (Astington & Jenkins, 1995) and deception (Ding, Wellman, Wang, Fu, & Lee, 2015). Once children have access to an explicit representational theory of mind, they become able to incorporate information about false beliefs (and other types of mental states) into their explicit judgments and evaluations of behavior and deliberative reasoning processes (Chalik, Rivera, & Rhodes, 2014; Rhodes & Brandone, 2014). Thus, children become more likely to consider information about mental states in their moral reasoning as they develop an explicit representational theory of mind. We discuss the relation of the present proposal to infancy research in more detail below.

There is substantial evidence indicating that the development of explicit theories of mind underlies developmental changes in moral judgment. It has long been shown that children increasingly incorporate information about mental states into their moral judgments across childhood. For example, Piaget (1965/1932) demonstrated that younger children primarily attended to outcome information over intentions to evaluate actions—they judged someone who accidentally made a large spill as worse than someone who intentionally made a small one, whereas older children and adults base such judgments on intent information (see Armsby, 1971; Baird & Astington, 2004; Cushman, Sheketoff, Wharton, & Carey, 2013; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011).

More recent work has shown that younger children, and even infants (Hamlin, 2013a, 2013b; Hamlin & Baron, 2014), can incorporate intent information into their moral evaluations to some extent. Nonetheless, the use of intent (over outcome) information becomes more robust across the childhood years and increased use of intent over outcome is predicted by developments in children's theories of mind (Cushman et al., 2013; Killen et al., 2011). For example, at earlier stages of development of children's intuitive theories, they are more likely to blame agents for their accidental transgressions (when the agents actually lacked the requisite mental states) and to excuse people who act with malicious intent but fail to cause harm (Cushman et al., 2013). While much of the research relating changes in intuitive psychology to moral development has been correlational (for example that changes in intuitive psychology correlate with changes in children's moral judgments, Smetana, Jambon, Conry-Murray, & Sturge-Apple, 2012; and even with moral behaviors such as lying, Lee, 2013), Ding et al. (2015) used experimental methods similar to Rhodes and Wellman (2013) to manipulate children's theories of mind, and found that children experimentally induced to develop false belief understanding became more likely to engage in lying behavior—confirming more directly one way in which the development of intuitive psychology has broad implications for moral thought and action.

1.2. Obligations

Moral judgment entails evaluating whether someone has broken a particular type of *obligation.* Whereas early perspectives on moral development such as Piaget (1965/1932) suggested that children view all obligations as set by authority figures (e.g., parents or teachers), and therefore as changeable by those authority figures and to-be-followed mainly to avoid punishments, the last several decades of research has thoroughly demonstrated that even by the early preschool years (at least), children view some norms—particularly those related to preventing harm and maintaining fairness—as intrinsic, universal, and inflexible obligations (Smetana, 1981, 2006; Turiel, 1983). Young children predict that people will behave in line with these obligations (that they will behave pro-socially and avoid harming, Boseovski, 2006), explain behavior by appeal to such obligations (Rhodes, 2014; Smetana, 1981; Turiel, 1983), and evaluate transgressions of these norms as particularly problematic in contrast to transgressions of non-moral social conventions (Smetana, 1981; Turiel, 1983).

How are concepts of obligation represented and reasoned about within children's intuitive theories of the world? As described by Wellman and Miller (2008), concepts of obligations and permissions play integral roles in early intuitive psychologies (and continue to do so across development). Wellman and Miller note that obligations and permissions are themselves intentional-psychological aspects of the world, and therefore part of the scope of theory of mind. Indeed, in children's predictions (Kalish, 1998) and explanations (Hickling & Wellman, 2001) of obligation-relevant behaviors, children integrate information about obligations with information about relevant mental states (e.g., whether the agent knew about the obligation and intended to follow it). Conversely, children (Leslie, Knobe, & Cohen, 2006) and adults (Knobe, 2003, 2004) incorporate information about whether an obligation has been broken or not into their judgments of intentionality. Thus, in an important sense, concepts of *obligation* are central to intuitive psychology, in that young children (and adults) recognize obligations as motivators of individual action, and predict, evaluate, and explain behavior by integrating information about obligations with information about relevant mental states.

Yet, there is also a key sense in which children's understanding of obligations rests on their intuitive *sociology*. Intuitive sociology entails abstract expectations about how people relate to one another (Rhodes, 2012). Recent research on intuitive sociology has revealed that children's notions of obligation are importantly embedded in their representations of social groups (Kalish & Lawson, 2008; Rhodes, 2012; Wellman & Miller, 2008), such that beliefs about whether a person is bound by and will act according to a particular norm depend not only on information about the person's mental states (e.g., whether they knew about the obligation and desired to follow it), but also by the agent's membership in a particular social group. By at least age 4, children expect that the content of some moral obligations varies by group memberships (Kalish & Lawson, 2008) and perhaps even more fundamentally, that whether a person holds an obligation to another individual depends on whether the two are members of the same or different social groupings.

Even quite young infants expect social behavior to depend on an agent's group membership (Powell & Spelke, 2013) and social allegiances (Hamlin, Mahajan, Liberman, & Wynn, 2013; Liberman, Kinzler, & Woodward, 2014). For example, Rhodes, Hetherington, Brink, and Wellman (2015) showed 16-month-old infants scenarios involving two cooperative partnerships—for simplicity here, a pair of two dogs that interacted cooperatively to accomplish a goal (get a ball out of a box) and a pair of two cats that interacted cooperatively to accomplish the goal. Infants then saw scenes in which one of the dogs thwarted one of the cat's attempts to achieve the same goal (for example, stomped on the lid to keep the ball inside). Did infants form any expectations about how the other members of the pairs, who had not yet interacted with one another, would interact? Indeed, 16-month-olds looked longer (indicating a violation of expectation) when the target members of the pairs (i.e., the other dog and cat, who had never before interacted with one another) interacted cooperatively with each other instead of coming into conflict. Follow-up studies confirmed that infants only formed these expectations after seeing the initial instances of conflict between one member from each pair, and that they formed team-based expectations even when all of the actors were members of the same species (e.g., all dogs), provided there were sufficient cues for them to track the cooperative pairings in the first place. Thus, some expectations that *allegiances* constrain morally-relevant social interactions (hindering and helping, see Hamlin) are present even in infancy.

By preschool, children's beliefs about obligations are still more clearly embedded in their explicit beliefs about group-based social structure. At least by age 4, children expect social groups to be characterized by distinct obligations (e.g., relating to dress, foods, and other customs, Kalish & Lawson, 2008) and expect obligations to differ more dramatically across different groups than other types of properties (such as beliefs or preferences; Kalish, 2012). Further, and perhaps speaking to obligations in their most basic form, at least by age 3, children view people as holding special intrinsic, interpersonal obligations to their own group members. Evidence for this claim comes from studies showing that, between the ages of 3–8, children view agents as acting to avoid harming their own group members (Rhodes, 2012), evaluate moral transgressions more harshly when they occur among members of the same group than among members of different groups (Rhodes & Chalik, 2013), and are more likely to hold agents responsible for transgressions committed towards members of the agent's own group (Rhodes, 2014).

As one example, Rhodes (2012) found that by age 3 children reliably predicted (over 75% of the time) that agents would direct harm towards members of other groups rather than members of their own. This pattern was quite robust across childhood, and is particularly striking for two reasons. First, the categories tested in this work were novel and minimal—marked by shirt color (e.g., red shirts and blue shirts) and made-up labels (e.g., Flurps and Zazzes). Second, children were not themselves members of either group (e.g., they made predictions about third party actions, involving groups that were not personally relevant). Because the groups were novel, children could not rely on direct previous experiences (e.g., observations of girls more often harming boys than other girls, for example) and because the children were not themselves members of either group, they could not rely on their own generalized preferences for in-group members to answer these questions. Instead, their predictions reflect abstract expectations about how categories constrain social interactions, that is, that people—in general—are more likely to harm members of other groups, instead of members of their own.

In this work (Chalik & Rhodes, 2014; Rhodes, 2012), reliable and robust predictions of inter-group harm were present by age 3—that is, children predicted Flurps would readily harm Zazzes, instead of other Flurps. Nonetheless, children did not reliably predict that *pro-social* interactions—helping someone else—would be constrained by group membership until age 6. That is, at ages 3–5, children predicted that Flurps would help other Flurps and Zazzes equally often. This pattern led to the hypothesis that children's patterns of inferences regarding harmful and pro-social behaviors reflect their beliefs about the causal mechanisms that lead categories to constrain these types of social interactions—namely, that children view these behaviors as reflecting patterns of *obligations.* Drawing on work in moral philosophy, obligations not to harm reflect the most basic of our interpersonal obligations (Cushman, Gray, Gaffey, & Mendes, 2012; Haidt & Joseph, 2004; Knobe, 2003; Rai & Fiske, 2011; Shweder, Mahapatra, & Miller, 1990; Wainryb, 2006). Pro-social, helpful behaviors—while valuable and perhaps morally praiseworthy—may not be *obligated* in this same manner. For example, although children at some young age might view people as obligated *not* to steal cookies, they might think it is *nice*, but not obligatory, to share a cookie. On this account, it is only those behaviors that children construe as obligatory that are shaped by their representations of social groups. Thus, children view people as obligated not to harm members of their own group, and because this is about an *obligation*, they do not extend this notion across group boundaries. In contrast, they fail to—at least at early ages—view pro-social actions as falling under the same scope of obligation, and thus, do not make group-based predictions about these types of behavior.

To test this account—that the pattern of children's predictions reflects beliefs that people are intrinsically obligated not to harm only members of their own groups—Rhodes and Chalik (2013) examined children's moral evaluations of instances of inter-group and intra-group harm. Borrowing methods from Social Domain Theory (Smetana, 1981), they tested whether children's evaluations of how bad it is to harm depend on the presence of extrinsic rules—if they view a prohibition (e.g., not to hit) as stemming from an intrinsic obligation (not to harm), then it shouldn't matter if there is not a rule in place that prohibits hitting, people shouldn't do it anyway (they are intrinsically obligated not to harm). Thus, Rhodes and Chalik tested if children thought in this way more for harm that occurred among members of the same group than among members of different groups.

Indeed, although children initially responded that it was just as bad for a Flurp to tease a Zaz as for a Flurp to tease a Flurp, they thought that it was markedly less bad for a Flurp to tease a Zaz if there was no rule in place prohibiting teasing. The information about the rule had no effect on whether they thought it was wrong for a Flurp to tease and hurt a Flurp-- regardless of the explicit rules, it was wrong for an agent to harm a member of the agent's own group. In short, these data support the view that young children see people as intrinsically obligated not to harm members of their own groups, but do not see these obligations as extending across group boundaries.

Further evidence for this interpretation comes from studies of children's explanations for moral transgressions. The background reasoning is that if a person violates an *intrinsic* obligation, that reflects badly on that agent. So, if children view people as holding intrinsic obligations only to their own group members, then it is specifically violations of these obligations that reflect badly on the agent. In contrast, instances where people harm members of *other* groups would not be viewed as reflecting moral shortcomings of the agent in particular, but instead could reflect other aspects of the situation. As predicted from this reasoning, Rhodes (2014) found that 4-year-old children more often referenced the agent (e.g., "he's mean") to explain harm among members of the same group than among members of different groups. Conversely, they more often referenced the relationship between the two people (e.g., "they're enemies," "they're different") to explain harm among members of different groups. Children view agents as more morally culpable for harm that they commit towards members of their own group than towards members of other groups, further supporting the claim that understandings of group memberships shape children's sense of moral obligations, and thus crucially shape children's moral judgments and moral learning.

To this point we've outlined how, by preschool age, many children appear to have a theory that people hold intrinsic obligations not to harm members of their own groups. This developing theory shapes children's predictions, explanations, and evaluations of morally-relevant action. But this theory develops and changes in several ways across childhood, and in particular, it changes in response to evidence children receive, just as outlined earlier in our description of intuitive theory change. First, as indicated by the data in Rhodes (2012), children view a broader range of behaviors—e.g., helping as well as avoiding harm--as falling under the scope of agents' group-based obligations as they get older. In that work, at around age six, children began to reliably predict that agents will preferentially direct helpful actions towards members of their own groups. These data are consistent with the developmental proposal that children first view people as obligated not to harm but only later as also obligated to help and support members of their own group.

Chalik and Rhodes (2014) revealed one type of evidence that might prompt children to expand their beliefs about group-based obligation in this manner. Chalik and Rhodes (2014) examined parental explanations for morally relevant actions including their explanations of actions among members of the same and different groups. To do this, they asked parents and children to go through a picture book that showed, on different pages, members of the same group (e.g., Flurps with Flurps) or different groups (e.g., Flurps with Zazzes) and asked them to judge whether agents should engage in pro-social actions (e.g., "Should this Flurp share a cookie with this Flurp?") as well as anti-social actions (e.g., "Should this Flurp steal a cookie from this Zaz?"), and then to explain "Why or why not?" Parents overwhelmingly told their children that the agents should engage in the pro-social behaviors and should refrain from the anti-social actions, and for these basic proscriptions there were no differences according to whether the interactions involved members of the same or different groups.

Yet, there were differences—subtle but powerful differences—in how parents explained these decisions to their children. In particular, parents more often gave explicitly *moral, obligation-enforcing* explanations for interactions among members of the same group. For example, when explaining why a Flurp should share with another Flurp, they said, "Because it is important to be fair. It would be right to share one with him." In contrast, parents less often referred to moral concepts like "fairness" or whether something is "right" when they discussed why an agent should share with a member of another group (why a Flurp should share with a Zaz). Instead they said things like, "That would be a nice thing to do." In this way, parents subtly communicated that agents are obligated to uphold certain moral standards in their interactions with members of their own groups, but not so obligated in their interactions with members of other groups. Extending the same behaviors to members of other groups is nice (possibly valuable), but not obligated in the same manner. This is a potentially powerful difference in messages, but it is important to reinforce how subtle this effect is: It is not the case that parents explicitly referenced the group memberships; for example, they did not say, "He has to share and be fair to him because they are both Flurps." Instead, they were simply more likely to refer to moral concepts (e.g., fairness) when discussing interactions that occurred among members of the same groups but not other groups. Moral tuition, and arguably moral learning, includes such subtle information, scaffolded by framework notions of obligations and of groups.

1.3. Interactions between intuitive psychology and sociology

Thus far, we have proposed that moral judgment rests on the interaction of two framework theories of the world—that children rely on their intuitive sociology to determine if a person has violated a moral obligation, and then on their intuitive psychology to determine if the person is culpable for these actions. Thus, moral development and learning can come from changes in, and result in changes of, both of these underlying theories. From this perspective, as children begin to consider a broader range of mental states in their use of intuitive psychology to predict behavior, they will show increasing sensitivity to these mental states in their moral evaluation. Further, as children develop, broaden, or refine their beliefs about what "counts" as an obligation, they will treat different behaviors as reflecting moral transgressions.

To further illustrate how intuitive psychology and sociology can interact with one another to underlie moral judgment across development, consider research by Chalik et al. (2014). In this research, children aged 3–5 were again introduced to "Flurps" and "Zazzes" and were asked to predict towards whom an agent would direct a harmful interaction, but this time were given conflicting information about social categories and about mental states. For example, would a Flurp hit another Flurp that he happened to be angry with, or a Zaz with whom he was not personally angry? Control conditions showed that in the absence of information about mental states, children expected Flurps to harm Zazzes rather than other Flurps (for example), and that in the absence of information about groups, they expected agents to harm people with whom they were angry, rather than harm people towards whom they felt neutrally. So far these data mimic those in other studies. The key question, however, was whether children would override their generalized expectations about how group members should and do behave, in the presence of conflicting information about an individual agent's mental states. Indeed children did so, but only once they had developed full-fledged representational theories of mind. In particular, only children who passed explicit false beliefs tasks reliably based their inferences on mental states (predicting that a Flurp would hit another Flurp that he was angry with, rather than a Zaz towards whom he felt neutrally). Children who did not hold fully representational theories of mind did not. Thus, before children had more fully developed theories of mind, they relied more strongly on their group-based intuitions.

These findings show developmental change not only in children's intuitive psychological and sociological theories, but also in how children integrate information from these two sources across childhood. In particular, children's beliefs that groups shape moral obligations appears robust already by age 3 (Rhodes, 2012; Rhodes & Chalik, 2013), and perhaps earlier in infancy (Rhodes et al., 2015), whereas several relevant aspects of intuitive psychology take considerably longer to develop. Thus, young children (in particular) might strongly rely on group information, with information about individual mental states playing an increasingly stronger role across development. This kind of shift is consistent with a broad range of findings in how children integrate information about categories and individuals to predict human action across multiple different tasks. For example, Diesendruck and HaLevi (2006) found that preschool-age children predicted behavior by attending to social category memberships instead of personality traits (see also Berndt & Heller, 1986; Biernat, 1991; Taylor, Rhodes, & Gelman, 2009) and Kalish and Shiverick (2004) found that preschool-age children predicted that people will follow their obligations even when they conflict with personal preferences, whereas older children expect personal preferences to more strongly influence individual behavior.

1.4. Further consideration of recent infant research

How might the proposed account, which emphasizes the processes that underlie developmental change, be reconciled with the now robust evidence of the sophisticated representational and inferential abilities of infants? Although we have argued that the *explicit* theories of mind that children develop in the preschool years are particularly important for moral judgment and behavior—in contrast to the more implicit understandings evidenced by infants—we do not view these implicit and explicit representations are wholly distinct, as implied by strict versions of dual-processing theories (Apperly, 2011). Instead, early social cognitive understandings and inferential abilities importantly feed into children's later more explicit understandings. Thus, individual differences in infants' attention to intentional action at 10–12 months predict later explicit theory of mind understanding at age 4, even controlling for individual differences in general processing abilities (Wellman, Lopez-Duran, LaBounty, & Hamilton, 2008, see also Brink, Lane, & Wellman, 2015; Thoermer, Sodian, Vuori, Perst, & Kristen, 2012; Yamaguchi, Kuhlmeier, Wynn, & VanMarle, 2009). From our perspective, early infant representations guide attention to new evidence in a manner that facilitates later conceptual change—in this way, early social cognitive understandings contribute to the development of children's later, more explicit representations (see detailed discussion in Wellman, 2014).

Further, from our perspective, the development of infants' early, more implicit, theories of mind can also be explained by the same processes that we have outlined here. Evidence of infant knowledge is often taken as support for nativist accounts—that whatever knowledge being tested is not learned through experience (see Spelke & Kinzler, 2009). Thus, infant ability to track individuals' mental states in infant "false belief" tasks is typically taken as evidence that such early implicit theory-of-mind representations are innate (Baillargeon, Scott, & He, 2010; Leslie, 1994). However, quite young infants possess powerful learning and inferential abilities that enable them to develop abstract knowledge even within the first year of life. Demonstrations of infant statistical learning famously show just this (e.g., Aslin, Saffran, & Newport, 1998; and for a direct demonstration of this sort of abstract learning from statistical data with regard to mental states, see Wellman, Kushnir, Xu, & Brink, 2016). Indeed, recent evidence suggests that the development of implicit theories of mind in infancy is critically dependent on experience. For example, whereas typically developing infants routinely track false beliefs in anticipatory looking paradigms at 18- to 25- months (Southgate, Senju, & Csibra, 2007), deaf infants of hearing parents (who are exposed to less language regarding internal mental states, see Meins et al., 2002) do not (Meristo et al., 2012). This pattern highlights the important role that experience and exposure to evidence play in the development of even the more implicit theories held by infants, in contrast to more strictly nativisit or maturational accounts of early conceptual knowledge.

Finally, recent infant work is also consistent with our description of moral judgment as resting on the interplay of intuitive psychology and sociology, rather than as the purview of a single encapsulated domain. For example, recent work on moral judgment in infancy has found that infants generally prefer agents who help over those who harm (Hamlin, 2012; Hamlin, Wynn, & Bloom, 2007), indicating that they incorporate morally-relevant information into their social evaluations within the first year of life.

Yet, even quite young infants do so in a manner that also incorporates information about both the agent's mental states (Hamlin, 2013a) and the agent's social relationships (Hamlin, 2012). For example, infants prefer agents who help only when all of the relevant parties have access to the information they would need to have in order to help intentionally and knowingly (Hamlin, 2013b). Further, infants' general preference for those who help is reduced when they see an agent help social partners not of their own group (Hamlin, 2012). Thus, within the first year of life, infants' morally-relevant social evaluations do not reflect the operations of a distinct, encapsulated domain, but rather involve the interplay of their intuitive psychological and sociological beliefs (see Wellman & Miller, 2008).

1.5. Conclusions and implications

We have put forth a constructivist approach towards moral learning, in that we propose that children actively build intuitive theories of the world, actively revise these theories in response to new evidence, and that these intuitive theories underlie moral judgment. While similar in spirit, our approach diverges in important ways from earlier constructivist approaches (Kohlberg, 1971; Piaget, 1932), in that we describe developmental change as a process of gradual theory revision, instead of involving a succession of discrete stages. Further, although we suggest that children revise their intuitive theories based on experience, we take a broad view on what these experiences might entail. The "evidence" that could prompt theory-revision might include children's own observations and direct experiences of morally relevant action, but can also involve input from other sources, including rather subtle features of language. Further, the hierarchical structure of children's knowledge, as we have described it here, equips them with powerful inferential capabilities. Thus, children can hold highly abstract, coherent and generalized expectations that allow them to predict, explain, and evaluate wide ranges of specific morally relevant actions, including those with which they have no direct experience.

In several ways, our position is different than others that have typically dominated discussions of morality and moral development (see also Wellman & Miller, 2008). As just described in regard to infant research, one dominant alternative is nativism. Nativism is embodied in different ways in accounts such as those by Spelke and Kinzler (2007), Haidt (2012) and others (e.g., Hauser, 2006; Mikhail, 2007; Rottman & Young, 2015). To reiterate, nativist accounts privilege evolved, innate representations (Hauser, 2006; Mikhail, 2007) and intuitions (Haidt, 2012) that do not develop, but are instead highly constrained. Still from this vantage point, moral learning occurs. In one account of moral learning from a nativist stance, learning can involve up-regulating or down-regulating the extent to which particular moral concerns factor into moral judgment, depending on the moral values of one's community. For example, "moral tuning" perspectives (see Haidt, 2012; Rottman & Young, 2015) suggest that children are born with a limited number of specific moral concerns (e.g., regarding harm, fairness, loyalty, purity, respect for authority, and so on) and learning involves "tuning" each moral dial up or down, depending on how much each concern is valued in one's culture. From this perspective, moral judgment rests on its own domain (Hauser, 2006; Mikhail, 2007) or domains (Haidt, 2012), which operate separately from other aspects of children's psychological or sociological knowledge.

Although our view does not preclude the possibility of important innate representations, our perspective differs from these nativist perspectives in several ways. First, we attribute a more active role to processes within the mind of the child as contributing to developmental change. Moral tuning perspectives often describe the child's mind as very passive (see e.g., Rottman & Young, 2015)—equipped with innate knowledge, regulated in response to the environment, but without considering (or even rejecting the possibility of) active cognitive processes in a child's mind for integrating cultural input and experiences with their representations to bring about cognitive change. Second, our approach can allow for more change, even qualitative change, and a less constrained system of moral judgment than described by most nativist perspectives. Finally, and perhaps most importantly, we do not view moral thinking as necessarily involving its own highly distinct domain (or domains), but rather as resting on the inferential capabilities afforded by children's more general intuitive theories of the psychological and social worlds.

In this way our position also differs from that of Social Domain Theory (Turiel, 1983). Although we share the broad view expressed by Social Domain Theory that children actively construct domain-specific theories of the world that underlie social and moral evaluation, our view also departs from this perspective in several respects. First, we do not define the domains in the same manner as researchers from this perspective. Social Domain Theory suggests that whether children construe an action as "moral" depends on its content (e.g., whether an action causes harm vs. violates a conventional norm such as manner of dress). From this perspective, children represent concerns about group membership as *conventional*, not moral. On this account, children identify moral infractions (e.g., someone caused harm by taking another's snack without asking), separately consider any group-relevant conventional concerns (e.g., did the person's action also violate or maintain a convention of group loyalty), and then weigh or integrate these concerns to form a final judgment (e.g., they might judge an action as wrong, and give a moral reason or this judgment, or as right, and appeal to conventional, group-based concerns; for review see Killen, 2007; Rutland, Killen, & Abrams, 2010). In contrast, from our perspective, group membership will often constrain whether an action will be identified as an infraction in the first place, because children view obligations as bounded by the group memberships of the agent and recipient, as we have reviewed above.

A further difference from Social Domain Theory (as well as other, older constructivist approaches) is that although we describe children's representations as *theories*, we further describe them as *intuitive theories*, in that we do not require that children can always fully articulate their beliefs and or that children will always have conscious access to the calculations that underlie their moral judgments. While our perspective does not preclude the possibility that *some* of moral judgment does indeed depend on conscious deliberation, we view our intuitive theories perspective as compatible with the large body of work documenting the role of intuition in moral judgment (and in particular, documenting that people do not always have conscious access to the reasons for why they make the judgments that they do, Cushman, Young, & Hauser, 2006). Although examining reasoning and justifications has a long tradition in research on moral thinking (Kohlberg, 1971; Smetana, 1981; Turiel, 1983), and is a valuable approach in some instances, our view of "theories" is consistent with the perspective that sometimes such theories are rather difficult to articulate (certainly this is the case of the theories held by preverbal babies, Gopnik & Meltzoff, 1997; Hamlin, 2013a, 2013b; Rhodes et al., 2015, as just discussed in our focused appraisal of infant moral and ToM research). Finally, as described earlier, we take a broad view of the types of evidence that can shape children's learning, including multiple features of language, observed experiences, as well as direct observations of the consequences of their own actions. In contrast, Social Domain Theory has tended to focus on children's direct experiences as the main instigator of moral learning and change.

Another dominant account privileges change but does so in terms of stage-like constructivism as described by Piaget and

Kohlberg. From the perspective of Piaget and Kohlberg, moral learning involves a slow constructive process across childhood, wherein children pass through a series of qualitatively distinct stages that are tied to domain general changes in children's cognitive capacities (e.g., abilities for abstract thinking and hypothetical reasoning). Clearly in this position moral learning is possible, indeed rampant, but it has a distinctive and problematic character (Haidt, 2012, provides a critique of this sort moral constructivism). Further, this approach clearly underestimates some of the moral reasoning capabilities of young children, as shown by decades of work from the perspective of Social Domain Theory and more recent infant research (e.g. Hamlin, 2012).

The position we advocate is neither of these, and fortunately the relevant terrain is not so dichotomous or limited. We propose instead that learning is constrained by deep abstract representations, but those representations can be (and often are) acquired via theory-based evidential learning. Moreover, they can be revised developmentally on the basis of theory-based evidential learning, which can then frame further learning. Intuitive psychology and sociology are not, from this perspective, intuitive because they are innate. Instead, they are intuitive because, once acquired, they frame human thinking and learning about morality. Further, although intuitive theories may not be explicitly articulated in the manner of scientific theories, they nevertheless function as theories, in that they shape how people predict, explain, and evaluate events in their environment. Importantly, with regard to human morality and moral learning, the mechanisms that propel development forward include learning within but also in the necessary interplay between intuitive sociology and intuitive psychology. Moral learning is important and complicated, as this special issue demonstrates. Tracking development is crucial to sorting out and understanding the complexity and importance of this fundamental feature of the human mind.

The field of moral psychology has documented numerous fascinating instances of developmental change—for example, that information about *intentions* play a greater role in moral judgment as children grow older (as discussed above), that judgments of fairness shift from concerns about strict equality to concerns about merit (Damon, 1975; Hook & Cook, 1979), that gaps between moral knowledge and moral behavior decrease with age (Shaw et al., 2014; Smith, Blake, & Harris, 2013), and that children become more concerned with deeper inequities across age (Blake & McAuliffe, 2011). Here we provide a framework for understanding the cognitive mechanisms underlying these developmental changes. Our proposal points to the need for more process-oriented developmental research, to examine how children's intuitive theories of the psychological and sociological worlds change in response to new evidence in a manner that instigates these changes. Thus, while the perspective that we put forth here is not meant to account for all of moral psychology, we do aim to describe a particular account of the learning mechanisms that underlie developmental change in moral thought.

References

Apperly, I. (2011). *Mindreaders: The cognitive basis of "theory of mind"*. New York: Psychology Press.

Armsby, R. E. (1971). A reexamination of the development of moral judgments in children. *Child Development, 42*(4), 1241–1248. http://dx.doi.org/10.2307/1127807.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324. http://dx.doi.org/10.1111/1467-9280.00063.

Astington, J. W., & Jenkins, J. M. (1995). Theory of mind development and social understanding. *Cognition & Emotion, 9*(2–3), 151–165. http://dx.doi.org/10.1080/02699939508409006.

Astington, J. W. (2003). Sometimes necessary, never sufficient: False-belief understanding and social competence. In B. Repacholi, V. Slaughter, & Virginia (Eds.), *Macquarie monographs in cognitive science* (pp. 13–38). New York, NY: Psychology Press.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*(3), 110–118. http://dx.doi.org/10.1016/j.tics.2009.12.006.

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. http://dx.doi.org/10.1002/cd.96.

Berndt, T. J., & Heller, K. A. (1986). Gender stereotypes and social inferences: A developmental study. *Journal of Personality and Social Psychology, 50*, 889–898. http://dx.doi.org/10.1037/0022-3514.50.5.889.

Biernat, M. (1991). Gender stereotypes and the relationship between masculinity and femininity: A developmental analysis. *Journal of Personality and Social Psychology, 61*, 351–365. http://dx.doi.org/10.1037/0022-3514.61.3.351.

Blake, P. R., & McAuliffe, K. (2011). "I had so much it didn't seem fair": Eight-year-olds reject two forms of inequity. *Cognition, 120*, 215–224. http://dx.doi.org/10.1016/j.cognition.2011.04.006.

Boseovski, J. (2006). Children's use of frequency information for trait categorization and behavioral prediction. *Developmental Psychology, 42*, 500–513. http://dx.doi.org/10.1037/0012-1649.42.3.500.

Brink, K. A., Lane, J. D., & Wellman, H. M. (2015). Developmental pathways for social understanding: Linking social cognition to social contexts. *Frontiers in Psychology, 6*. http://dx.doi.org/10.3389/fpsyg.2015.00719.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: Bradford Books, MIT Press.

Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press.

Chalik, L., & Rhodes, M. (2014). Preschoolers use social allegiances to predict behavior. *Journal of Cognition and Development, 15*, 136–160. http://dx.doi.org/10.1080/15248372.2012.728546.

Chalik, L., Rivera, C., & Rhodes, M. (2014). Children's use of categories and mental states to predict social behavior. *Developmental Psychology, 50*(10), 2360. http://dx.doi.org/10.1037/a0037729.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analysis in moral judgment. *Cognition, 108*(2), 353–380. http://dx.doi.org/10.1016/j.cognition.2008.03.006.

Cushman, F. A., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion, 12*(1), 2–7. http://dx.doi.org/10.1037/a0025071.

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*(12), 1082–1089. http://dx.doi.org/10.1111/j.1467-9280.2006.01834.x.

Damon, W. (1975). Early conceptions of positive justice as related to the development of logical operations. *Child Development, 46*, 301–312. http://dx.doi.org/10.2307/1128122.

Diesendruck, G., & HaLevi, H. (2006). The role of language, appearance, and culture in children's social category-based induction. *Child Development, 77*, 539–553. http://dx.doi.org/10.1111/j.1467-8624.2006.00889.x.

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-mind training causes honest young children to lie. *Psychological Science, 26*(11), 1812–1821. http://dx.doi.org/10.1177/0956797615604628.

Dunn, J., Cutting, A. L., & Demetriou, H. (2000). Moral sensibility, understanding others, and children's friendship interactions in the preschool period. *British Journal of Developmental Psychology, 18*(2), 159–177. http://dx.doi.org/10.1348/026151000165625.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford, England: Oxford University Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind and Language, 7*, 145–171. http://dx.doi.org/10.1111/j.1468-0017.1992.tb00202.x.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*, 1085–1108. http://dx.doi.org/10.1037/a0028044.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. Hirschfeld & S. Gelman (Eds.), *Domain specificity in cognition and culture* (pp. 257–293). New York, NY: Cambridge University Press.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY: Pantheon/Random House.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus: Special Issue on Human Nature, 133*, 55–66. http://dx.doi.org/10.1162/0011526042365555.

Hamlin, J. K. (2012). A developmental perspective on the moral dyad. *Psychological Inquiry, 23*(2), 166–171. http://dx.doi.org/10.1080/1047840X.2012.670101.

Hamlin, J. K. (2013a). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition, 128*, 451–474. http://dx.doi.org/10.1016/j.cognition.2013.04.004.

Hamlin, J. K. (2013b). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science, 22*(3), 186–193. http://dx.doi.org/10.1177/0963721412470687.

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. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me = bad infants prefer those who harm dissimilar others. *Psychological Science, 24*, 589–594. http://dx.doi.org/10.1177/0956797612457785.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557–559. http://dx.doi.org/10.1038/nature06288.

Hart, H. L. A. (1968). Legal responsibility and excuses. In S. Hook (Ed.), *Determinism and freedom in the age of modern science* (pp. 99–116). New York: New York University Press.

Hauser, M. D. (2006). *Moral minds: The nature of right and wrong*. New York, NY: Harper.

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology, 37*(5), 668. http://dx.doi.org/10.1037/0012-1649.37.5.668.

Hirschfeld, L. A. (1996). *Race in the making*. Cambridge, MA: MIT Press.

Hook, J. G., & Cook, T. D. (1979). Equity theory and the cognitive ability of children. *Psychological Bulletin, 86*(3), 429. http://dx.doi.org/10.1037/0033-2909.86.3.429.

Inagaki, K., & Hatano, G. (2002). *Young children's naive thinking about the biological world*. New York, NY: Psychology Press.

Kalish, C. W. (1998). Reasons and causes: Children's understanding of conformity to social rules and physical laws. *Child Development, 69*(3), 706–720. http://dx.doi.org/10.1111/j.1467-8624.1998.tb06238.x.

Kalish, C. W. (2012). Generalizing norms and preferences within social categories and individuals. *Developmental Psychology, 48*(4), 1133. http://dx.doi.org/10.1037/a0026344.

Kalish, C. W., & Lawson, C. A. (2008). Development of social category representations: Early appreciation of roles and deontic relations. *Child Development, 79*, 577–593. http://dx.doi.org/10.1111/j.1467-8624.2008.01144.x.

Kalish, C. W., & Shiverick, S. M. (2004). Children's reasoning about norms and traits as motives for behavior. *Cognitive Development, 19*, 401–416. http://dx.doi.org/10.1016/j.cogdev.2004.05.004.

Killen, M. (2007). Children's social and moral reasoning about exclusion. *Current Directions in Psychological Science, 16*, 32–36. http://dx.doi.org/10.1111/j.1467-8721.2007.00470.x.

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition, 110*, 197–215. http://dx.doi.org/10.1016/j.cognition.2011.01.006.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*, 190–194. http://dx.doi.org/10.1111/1467-8284.00419.

Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis, 64*, 181–187. http://dx.doi.org/10.1111/j.1467-8284.2004.00481.x.

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences, 9*, 357–359. http://dx.doi.org/10.1016/j.tics.2005.06.011.

Kohlberg, L. (1971). From is to ought. In T. Mischel (Ed.), *Cognitive development and epistemology* (pp. 151–235). New York, NY: Academic Press.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology, 43*, 186–197. http://dx.doi.org/10.1037/0012-1649.43.1.186.

Lee, K. (2013). Little liars: Development of verbal deception in children. *Child Development Perspectives, 7*, 91–96.

Leslie, A. M. (1994). *ToMM, ToBy, and agency: Core architecture and domain specificity. Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). Cambridge, England: Cambridge University Press.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect theory of mind and moral judgment. *Psychological Science, 17*(5), 421–427. http://dx.doi.org/10.1111/j.1467-9280.2006.01722.x.

Liberman, Z., Kinzler, K. D., & Woodward, A. L. (2014). Friends or foes: Infants use shared evaluations to infer others' social relationships. *Journal of Experimental Psychology: General, 143*(3), 966. http://dx.doi.org/10.1037/a0034481.

Meins, E., Fernyhough, C., Wainwright, R., Das Gupta, M., Fradley, E., & Tuckey, M. (2002). Maternal mind–mindedness and attachment security as predictors of theory of mind understanding. *Child Development, 73*(6), 1715–1726. http://dx.doi.org/10.1111/1467-8624.00501.

Meristo, M., Morgan, G., Geraci, A., Iozzi, L., Hjelmquist, E., Surian, L., & Siegal, M. (2012). Belief attribution in deaf and hearing infants. *Developmental Science, 15* (5), 633–640. http://dx.doi.org/10.1111/j.1467-7687.2012.01155.x.

Mikhail, J. M. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152. http://dx.doi.org/10.1016/j.tics.2006.12.007.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*(3), 289. http://dx.doi.org/10.1037/0033-295X.92.3.289.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258. http://dx.doi.org/10.1126/science.1107621.

Peterson, C. C., & Siegal, M. (2002). Mindreading and moral awareness in popular and rejected preschoolers. *British Journal of Developmental Psychology, 20*(2), 205–224. http://dx.doi.org/10.1348/026151002166415.

Peterson, C. C., Slaughter, V. P., & Paynter, J. (2007). Social maturity and theory of mind in typically developing children and those on the autism spectrum. *Journal of Child Psychology and Psychiatry, 48*(12), 1243–1250. http://dx.doi.org/10.1111/j.1469-7610.2007.01810.x.

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

Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences, 110*(41), E3965–E3972. http://dx.doi.org/10.1073/pnas.1304326110.

Rai, T., & Fiske, A. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality and proportionality. *Psychological Review, 118*(1), 57–75. http://dx.doi.org/10.1037/a0021867.

Rhodes, M. (2012). Naive theories of social groups. *Child Development, 83*, 1900–1916. http://dx.doi.org/10.1111/j.1467-8624.2012.01835.x.

Rhodes, M. (2014). Children's explanations as a window into their intuitive theories of the social world. *Cognitive Science, 38*, 1687–1697. http://dx.doi.org/10.1111/cogs.12129.

Rhodes, M., & Brandone, A. (2014). Three-year-olds' theories of mind in actions and words. *Frontiers in Developmental Psychology, 5*, 1–8. http://dx.doi.org/10.3389/fpsyg.2014.00263.

Rhodes, M., & Chalik, L. (2013). Social categories as markers of intrinsic interpersonal obligations. *Psychological Science, 24*(6), 999–1006. http://dx.doi.org/10.1177/0956797612466267.

Rhodes, M., Hetherington, C., Brink, K., & Wellman, H. (2015). Infants' use of social partnerships to predict behavior. *Developmental Science, 18*, 909–916. http://dx.doi.org/10.1111/desc.12267.

Rhodes, M., & Wellman, H. (2013). Constructing a new theory from old ideas and new evidence. *Cognitive Science, 37*, 592–604. http://dx.doi.org/10.1111/cogs.12031.

Rottman, J., & Young, L. (2015). Mechanisms of moral development. In J. Decety & T. Wheatley (Eds.), *The moral brain: A multidisciplinary perspective* (pp. 123–142). Cambridge, MA: MIT Press.

Rutland, A., Killen, M., & Abrams, D. (2010). A new social-cognitive developmental perspective on prejudice the interplay between morality and group identity. *Perspectives on Psychological Science, 5*(3), 279–291. http://dx.doi.org/10.1177/1745691610369468.

Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development, 80*(4), 1172–1196. http://dx.doi.org/10.1111/j.1467-8624.2009.01324.x.

Scott, R. M., Baillargeon, R., Song, H. J., & Leslie, A. M. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology, 61*(4), 366–395. http://dx.doi.org/10.1016/j.cogpsych.2010.09.001.

Shaw, A., Montinari, N., Piovesan, M., Olson, K. R., Gino, F., & Norton, M. I. (2014). Children develop a veil of fairness. *Journal of Experimental Psychology: General, 143*(1), 363. http://dx.doi.org/10.1037/a0031247.

Shweder, R. A., Mahapatra, M., & Miller, J. G. (1990). Culture and moral development. In J. W. Stigler, R. A. Shweder, & G. Herdt (Eds.), *Cultural psychology: Essays on comparative human development* (pp. 130–204). Cambridge: Cambridge University Press.

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron, 41*(4), 653–662. http://dx.doi.org/10.1016/S0896-6273(04)00014-5.

Slaughter, V., Imuta, K., Peterson, C. C., & Henry, J. D. (2015). Meta-analysis of theory of mind and peer popularity in the preschool and early school years. *Child Development, 86*, 1159–1174. http://dx.doi.org/10.1111/cdev.12372.

Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child Development, 52*, 1333–1336. http://dx.doi.org/10.2307/1129527.

Smetana, J. G., Jambon, M., Conry-Murray, C., & Sturge-Apple, M. L. (2012). Reciprocal associations between young children's developing moral judgments and theory of mind. *Developmental Psychology, 48*, 1144–1155. http://dx.doi.org/10.1037/a0025891.

Smetana, J. (2006). Social domain theory: Consistencies and variations in children's moral and social judgments. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 119–154). Mahwah, NJ: Erlbaum.

Smith, C. E., Blake, P. R., & Harris, P. L. (2013). I should but I won't: Why young children endorse norms of fair sharing but do not follow them. *PLoS ONE, 8*(3), e59510. http://dx.doi.org/10.1371/journal.pone.0059510.

Song, H. J., & Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology, 44*(6), 1789. http://dx.doi.org/10.1037/a0013774.

Song, H. J., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition, 109*(3), 295–315. http://dx.doi.org/10.1016/j.cognition.2008.08.008.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592. http://dx.doi.org/10.1111/j.1467-9280.2007.01944.x.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science, 10*, 89–96. http://dx.doi.org/10.1111/j.1467-7687.2007.00569.x.

Spelke, E. S., & Kinzler, K. D. (2009). Innateness, learning and rationality. *Child Development Perspectives, 3*, 96–98. http://dx.doi.org/10.1111/j.1750-8606.2009.00085.x.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580–586. http://dx.doi.org/10.1111/j.1467-9280.2007.01943.x=.

Taylor, M., Rhodes, M., & Gelman, S. (2009). Boys will be boys; cows will be cows: Children's essentialist reasoning about gender categories and animal species. *Child Development, 80*, 461–481. http://dx.doi.org/10.1111/j.1467-8624.2009.01272.x.

Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology, 30*(1), 172–187. http://dx.doi.org/10.1111/j.2044-835X.2011.02067.x.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge: Cambridge University Press.

Wainryb, C. (2006). Moral development in culture: Diversity, tolerance, and justice. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 211–240). Mahwah, NJ: Erlbaum.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. New York, NY: Oxford University Press.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684. http://dx.doi.org/10.1111/1467-8624.00304.

Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory-of-mind scale: Longitudinal perspectives. *Child Development, 82*(3), 780–792. http://dx.doi.org/10.1111/j.1467-8624.2011.01583.x.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*, 337–375. http://dx.doi.org/10.1146/annurev.ps.43.020192.002005.

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016). Infants use statistical sampling to understand the psychological world. *Infancy*. http://dx.doi.org/10.1111/infa.1213.

Wellman, H. M., & Liu, D. (2004). Scaling of theory of mind tasks. *Child Development, 75*, 523–541. http://dx.doi.org/10.1111/j.1467-8624.2004.00691.x. *82*, 780–792. doi: 10.1111/j.1467-8624.2011.01583.x.

Wellman, H. M., Lopez-Duran, S., LaBounty, J., & Hamilton, B. (2008). Infant attention to intentional action predicts preschool theory of mind. *Developmental Psychology, 44*(2), 618. http://dx.doi.org/10.1037/0012-1649.44.2.618.

Wellman, H. M., & Miller, J. (2008). Including deontic reasoning as fundamental to theory of mind. *Human Development, 51*, 105–135. http://dx.doi.org/10.1159/000115958.

Xu, F. (2007). Rational statistical inference and cognitive development. In P. Carruthers, S. Laurence, & S. Stich (Eds.). *The innate mind: Foundations and the future* (Vol. 3, pp. 199–215). New York, NY: Oxford University Press.

Yamaguchi, M., Kuhlmeier, V. A., Wynn, K., & VanMarle, K. (2009). Continuity in social cognition from infancy to childhood. *Developmental Science, 12*(5), 746–752. http://dx.doi.org/10.1111/j.1467-7687.2008.00813.x.

Young, L., Cushman, F., Hauser, M., & 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. http://dx.doi.org/10.1073/pnas.0701408104.