20

The Moral Baby

Karen Wynn and Paul Bloom

Most developmental research into morality so far has focused on children and adolescents, as can be seen in the contributions of this current volume. We think that the time is ripe to take a serious look at the moral lives of babies.

One motivation for this comes from evolutionary theory. Biologists have long been interested in how a species like ours—in which large groups of nonkin work together on projects of mutual benefit—could come to exist. This was largely a mystery at the time of Darwin, but there are by now several candidate theories for how our complex social structures can arise. These include the accounts developed in the 1970s and 1980s based on kin selection and reciprocal altruism (e.g., Axelrod, 1984; Trivers, 1971, 1985), as well as theories based on group selection—a proposal once derided by biologists, but now returning as a serious contender (see Nowak & Highfield, 2011, for an accessible review). Such theories explain our complex social structures as grounded in certain propensities that we can view as moral, including altruism to nonkin, guilt at betraying another, and righteous anger toward cheaters. While the details are a matter of considerable debate, the notion of unlearned moral universals is consistent with what we now know about biological evolution. And one way to explore the nature of such universals is to look at babies.

The second motivation comes from developmental psychology. Over the last 30 or so years, findings based on looking-time methods set off a revolution in how we think about the minds of babies. The original studies used such methods to focus on early knowledge of physical objects—a baby's "naïve physics." A vast body of research now suggests that—contrary to what was taught for decades to legions of psychology undergraduates—babies think of objects largely as adults do, as connected masses that move as units, that are solid and subject to gravity, and that move in continuous paths through space and time (e.g., Baillargeon, 1987; Spelke, 1990; Wynn, 1992). Other studies have found rich social understanding. For instance, babies before their first birthday appreciate that individuals have goals (Gergely et al., 1995; Woodward, 1998) and soon afterward they appreciate the other individuals can have false beliefs (Onishi & Baillargeon, 2005). These sorts of findings make it plausible that some rudimentary moral capacities will also be present in young babies.

In a sense, then, ours is a "nativist" approach, in that it takes seriously the proposal of an innate morality. But we do see this as an empirical proposition, to be supported or falsified by empirical study. Moreover, adults possess certain moral propensities and judgments that we should expect, from an evolutionary perspective, *not* to show up in young babies. These include aspects of morality that have no genetic payoff, such as kindness to distant strangers. We will return to this issue at the end of the chapter.

Some investigators, including many in this volume, begin their inquiry by proposing an explicit definition of morality. Certainly one needs to have a rough sense of what one means by morality in order to study it, but we think that starting with a definition is ill advised. After all, there is no agreed-upon definition of morality by moral philosophers (Nado, Kelly, & Stich, 2009), and psychologists sharply disagree about what is and is not moral (see, e.g., Turiel, 1998, and Haidt, 2012, for conflicting proposals.). Some argue that from a psychological perspective, morality does not correspond to a natural kind—that is, what we normally talk about as "moral" corresponds to distinct neural and cognitive systems (Sinnott-Armstrong & Wheatley, in press). In the end—just as with notions such as language and memory—the proper scope of morality (as a psychological construct) is an empirical question, to be resolved as we learn more about the mind.

Moral Sentiments

One can distinguish moral understanding from moral sentiments. It is one thing to judge that certain acts are right or wrong—to appreciate, for instance, that if X hits Y for no reason, then X has done something wrong. It is another to have moral emotions, to feel sympathy for the pain of Y and anger toward X. While most of the research that we discuss below focuses on understanding, there is little doubt that such sentiments are critical to morality. As David Hume (1739/2000) pointed out, without moral *passions*, our moral reasoning would be useless—we might know right from wrong, but we would never be motivated to act upon this knowledge.

There are several moral emotions, including guilt, shame, gratitude, and anger, but most developmental research has focused on caring about other people—sometimes described as compassion. Is this an inherent part of our natures?

Many scholars believe that it is, that it makes society and culture possible. In his book *The Theory of Moral Sentiments*, published in 1759, Hume's contemporary Adam Smith begins with:

> How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it. Of this kind is pity or compassion, the emotion which we feel for the misery of others, when we either see it, or are made to conceive it in a very lively manner. (Smith, 2002, p. 11)

We see hints of such moral "principles" early on. Laboratory studies find that young infants just a few months old will become distressed and cry when they hear the cry of another baby (Sagi & Hoffman, 1976), and this is supported anecdotally. In his classic article on the development of his son, William, Charles Darwin describes, as evidence of the

"moral emotions," his son's reaction to the suffering of others: "With respect to the allied feeling of sympathy, this was clearly shown at 6 months and 11 days by his melancholy face, with the corners of his mouth well depressed, when his nurse pretended to cry" (Darwin, 1877, p. 294).

Contemporary research finds that such responses are not merely reactions to aversive noise; babies do not cry as much when they hear a recording of their own cry (Dondi, Simion, & Caltran, 1999) or the cry of an infant chimpanzee (Martin & Clark, 1982). Still, a skeptic might wonder whether this crying really reflects compassion, as opposed to, perhaps, reflecting a competitive motivation (if another baby is claiming the attentions of an adult, their own crying might serve well to get in on the action themselves). We know that this isn't entirely the case for other creatures, as they engage in a range of behaviors to stop the pain of other members of their species. Hungry rhesus monkeys avoid pulling a lever to get food if a lever pull gives another monkey a painful electric shock (Wechkin, Masserman, & Terris, 1964; Masserman, Wechkin, & Terris, 1964). Rats will press a bar to lower another rat suspended in midair or trapped in a tank full of water; and, like monkeys, rats will refrain from pressing a bar that provides food if the bar press shocks another rat (Rice & Gainer, 1962; Rice, 1964).

What about human toddlers? There are many anecdotes of soothing actions by 1-year olds; the patting and soothing of others in distress (for review, see Hoffman, 2000). An influential series of studies by Carolyn Zahn-Waxler and her colleagues took this into the lab, analyzing how babies respond to acted-out pain by those around them, such as their mother banging her knee or an experimenter getting her finger caught in a clipboard (Zahn-Waxler et al., 1992a, 1992b). They found that toddlers often soothe in such situations (and they are more likely to do so for their parents than for strangers). Twin studies suggest that the extent of soothing is partially heritable (Zahn-Waxler et al., 1992b). Girls are more likely to soothe than boys, which meshes with a broader literature suggesting that females are more prone than males to empathy and altruism (see Baron-Cohen, 2004, for review.)

Other studies have found that toddlers will voluntarily act to assist an adult experimenter having difficulty with a task such as putting books away in a cupboard or hanging clothes on a line (Warneken & Tomasello, 2006; see also 2009), and will do so even at a cost to themselves (e.g., momentarily abandoning an especially enjoyable activity; Warneken & Tomasello, 2008). Interestingly, this behavior occurs not only without reward, but actually *decreases* under conditions of explicit reward and approval (Warneken & Tomasello, 2008), suggesting that helping others is intrinsically rewarding to the young human. Such findings suggest that young children have prosocial tendencies that influence their social actions and interactions; in some circumstances at least, they care about others, and this motivates certain positive actions.

A Moral Sense

Psychological research into the moral emotions of babies has been ongoing for a long time, but it's only recently that psychologists have explored the origins of what the philosophers of the Scottish Enlightenment described as a moral sense. This is not the same as an impulse to do good and to avoid doing evil. Rather, it is the capacity to make certain types

of judgments. Smith, though himself skeptical of its existence, provides the best description of the moral sense, as:

> somewhat analogous to the external senses. As the bodies around us, by affecting these in a certain manner, appear to possess the different qualities of sound, taste, odour, colour; so the various affections of the human mind, by touching this particular faculty in a certain manner, appear to possess the different qualities of amiable and odious, of virtuous and vicious, of right and wrong. (Smith, 2002, pp. 379–380)

Do babies have such a sense? At minimum, they are sensitive to the *valences* of different actions. Premack and Premack (1997) repeatedly presented 1-year olds with (*habituated* them to) a computer-animated display in which one character acted positively toward another character (by caressing it or by helping push it through a tight gateway it was trying to squeeze through); infants were shown this until they became bored as indicated by significantly decreased looking times. These infants looked significantly longer (*dishabituated*) when they were then shown displays in which one character acted negatively toward another by hitting it. In contrast, infants who were habituated to negative interactions (in which the first character hit the second, or prevented it from getting through the gateway by nudging it backward) did not dishabituate when subsequently shown an example of hitting—they continued to be bored by this new example of negative behavior. These results suggest that infants found a helping action to be similar to a caressing action, and a hindering action to be similar to an act of hitting. That is, preverbal infants recognized the commonality of valence shared by the two perceptually distinct prosocial interactions (helping and caressing), and by the two perceptually distinct antisocial interactions (hindering and hitting).

In our own studies, first conducted in collaboration with Kuhlmeier, we found that infants are sensitive to the valence of social interactions in predicting the behavior of others (e.g., Kuhlmeier, Wynn, & Bloom, 2003, under review; Hamlin, Wynn, & Bloom, 2007; see also Wynn, 2008). In these studies, infants were shown events in which a character—the Climber—repeatedly attempted to ascend a steep incline. In some trials, the Climber was helped uphill by an individual (the Helper) who nudged the Climber from behind; in others, the Climber was pushed downhill by a different individual (the Hinderer) (Figure 20.1A). Infants were then shown test events in which the Climber, on alternate trials, approached the Helper and approached the Hinderer (Figure 20.1B), and their looking times to these two events were measured. We found that infants discriminated these events in their looking times, suggesting that they expected the Climber to hold distinct attitudes toward the two characters. More specifically, when the characters had "faces," making them salient social beings, 9- and 10-month-old infants (but not 5- or 6-month olds) looked longer when the Climber approached the Hinderer than when it approached the Helper, suggesting that they expected the Climber to be inclined to avoid the Hinderer but not the Helper (Kuhlmeier, Wynn, & Bloom, under review; Hamlin, Wynn, & Bloom, 2007; see also Wynn, 2008).

Our interpretation of these findings is that they reflect a form of social evaluation, tapping infants' mental attributions to the characters. This view is supported by longitudinal evidence that individual infants' performance on this task at 1 year of age correlates

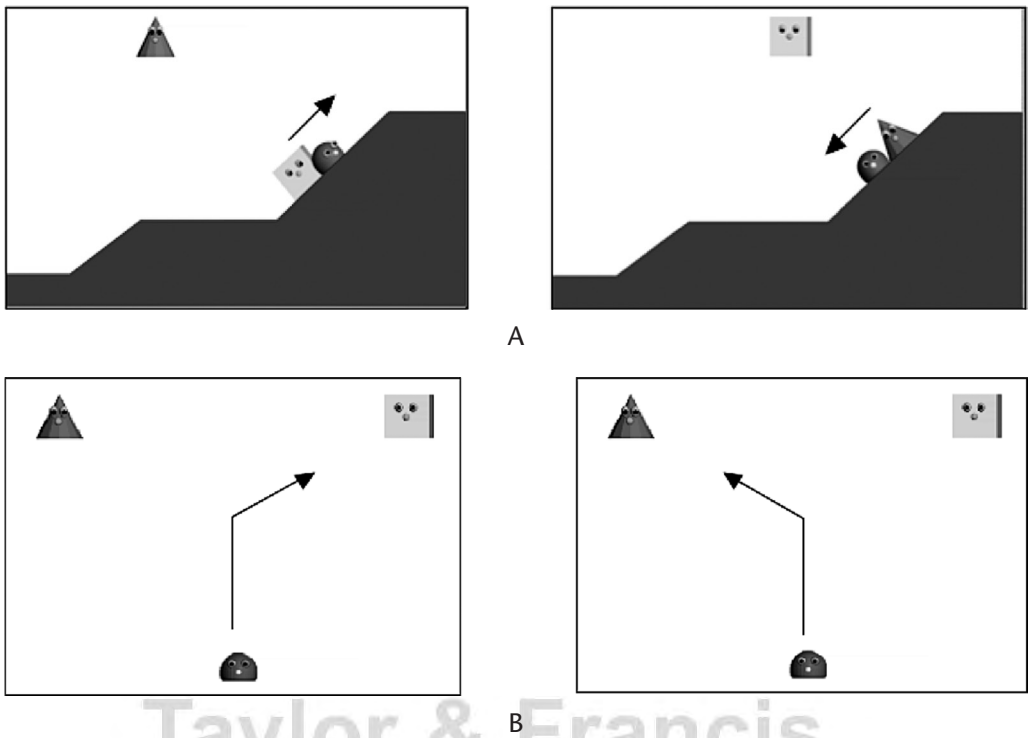


Figure 20.1 Computer-animated events of Kuhlmeier, Wynn, & Bloom (under review). See also Wynn, 2008. One character (in this example, the Square) helps the red Climber get up the hill by nudging it from behind; another (the Triangle) pushes it down the hill by nudging it from the front.

positively with their performance on a battery of theory of mind tasks at 4 years of age (Yamaguchi, Kuhlmeier, Wynn, & vanMarle, 2009). (In contrast, infants' performance on a nonsocial, auditory number-discrimination task (from vanMarle & Wynn, 2006) does not correlate with their later theory of mind performance.) Note that, in addition to supporting our interpretation that infants' responses in our "hill-climbing" task reflect their assessments of the mental states of others, such findings also suggest that there are stable individual differences in theory of mind capacities within the normal population, even from the first few months of life.

The studies described above assessed babies' capacity to generate social predictions; they did not probe for the possible presence of early moral evaluation. Accordingly, our next step was to ask how babies themselves *feel* about positive and negative social actions—and toward the actors who engage in them.

In collaboration with Hamlin, we have conducted a series of studies to address this issue. Our first study (Hamlin, Wynn, & Bloom, 2007) employed a scenario adapted from Kuhlmeier, Wynn, and Bloom (2003, under review). Infants witnessed a character—a painted wooden block with googly eyes glued onto it—repeatedly try to ascend a steep incline. On some attempts, a second character (another block with eyes, of a different shape and color from the Climber) helped the Climber up the hill by nudging it from

behind; on other attempts, a third character interfered with the Climber's efforts by pushing it downhill (see Figure 20.2, Panels A and B). Would infants see the former action as helpful and "good," and the latter action as unhelpful and "bad"? Would they feel warmly toward the helpful individual, and be negatively inclined toward the hindering one? Following habituation to these events, an experimenter (blind to which character was the "good" and which was "bad") held out both the helpful and the unhelpful characters to the infant simultaneously, and encouraged the infant to express its preference for one over the other by reaching for one of them (Figure 20.2, Panel C).

Figure 20.2 A. Helper pushes Climber up. B. Hinderer pushes Climber down. C. Whom does Baby prefer?

Credit (for A, B, and C): Live-presentation, three-dimensional "Hill" events enacted to infants in Hamlin, Wynn, & Bloom (2007; adapted from Kuhlmeier, Wynn, & Bloom, 2003, under review). One character (in this example, the Ball) helps the yellow Climber get up the hill by nudging it from behind; another (the Square) pushes it down the hill by nudging it from the front.

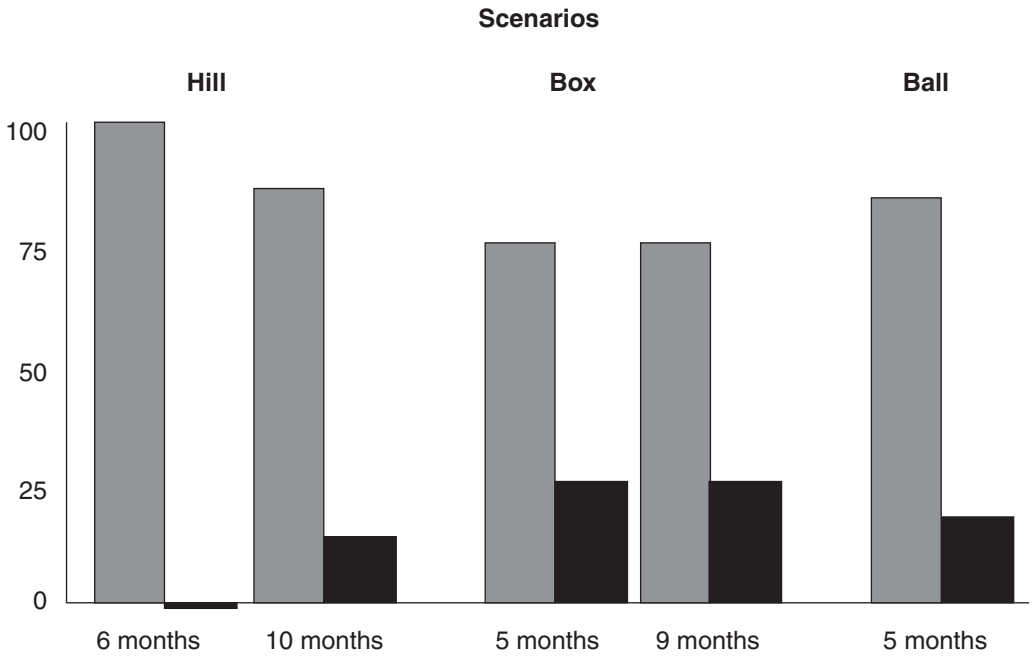


Figure 20.3 Scenarios

Credit: Percentage of infants of each age tested who chose the prosocial character (light bars) and percentage who chose the antisocial character (black bars), across three distinct social scenarios: "Hill" (Hamlin, Wynn, & Bloom 2007), "Box," and "Ball" (Hamlin & Wynn, 2011). Infants significantly preferred the prosocial character at all ages and in all scenarios, all p's < 0.05.

As predicted, both 6- and 10-month-old infants overwhelmingly preferred the helpful individual (see Figure 20.3, first four bars). Such a result can be explained three different ways, however: Babies might be drawn to the helpful individual; they might be repelled by the hindering individual; or both. We explored this question in a further series of studies that introduced a neutral character, one that neither helps nor hinders. We found that, given a choice, infants prefer a helpful character to a neutral one, and prefer a neutral character to one who hinders. This finding indicates that both inclinations are at work—babies are drawn to the nice guy and are repelled by the mean one. Again, these results were not subtle; babies almost always showed this pattern of response.

These findings have recently been extended to two additional social scenarios (Hamlin & Wynn, 2011). In the first, 5- and 9-month-old infants saw a puppet attempting to pry up the lid of a box, where the lid keeps falling back down after being raised partway. On alternating attempts, a "prosocial" puppet came forward and grasped the other side of the lid in the same manner as the protagonist, so helping to successfully open the box; and an "antisocial" puppet came forward and jumped onto the lid of the box, slamming it shut (see Figure 20.4). In the second new scenario, 5-month-old infants saw a puppet playing with a small ball. In alternating events, this puppet rolled the ball toward (a) a "prosocial" puppet, who picked it up and rolled it back to the protagonist, and (b) an "antisocial" puppet, who picked it up and absconded offstage with it (see Figure 20.5). As predicted, in both

A

B

Figure 20.4 A. Kitty on left helps puppy open box. B. Kitty on right jumps on box, closing it.

Credit (for A and B): "Box" events from Hamlin & Wynn (2011). Puppy attempts to open the box but cannot fully lift the lid. One kitty helps the puppy open the box; the other slams the lid shut.

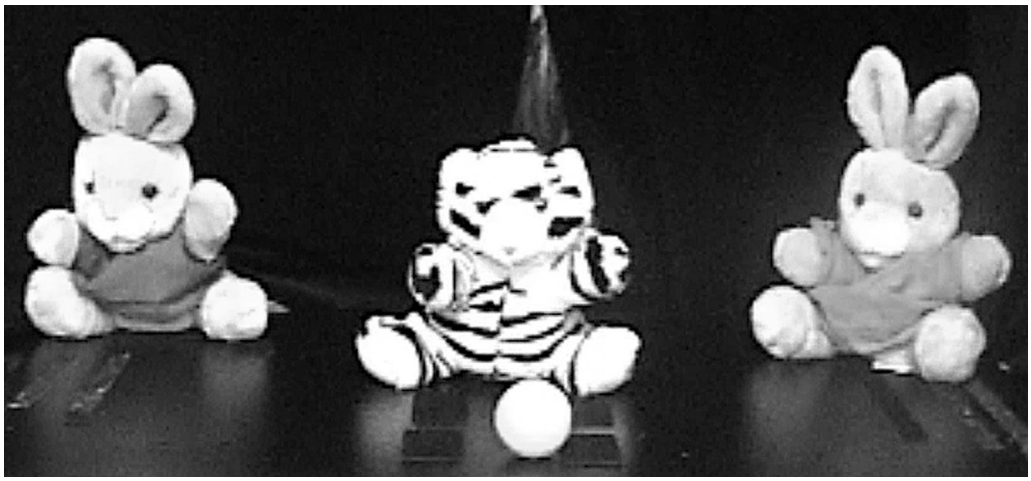


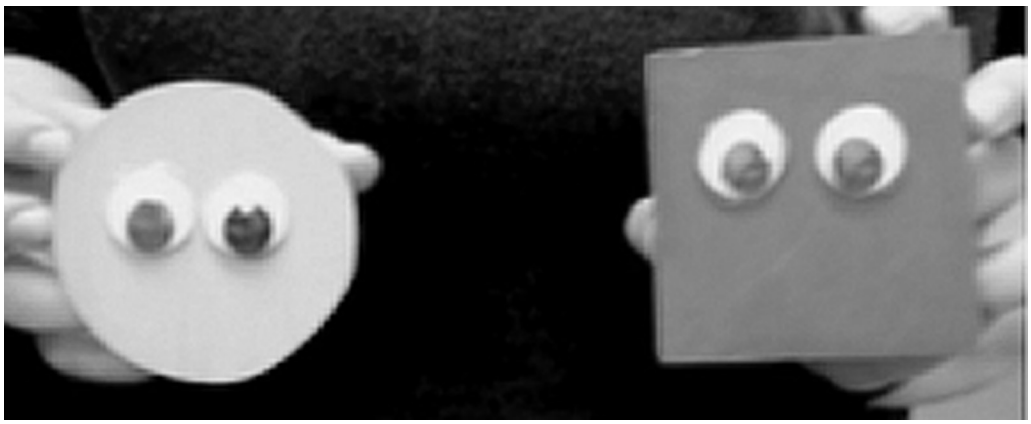
Figure 20.5 "Ball" events from Hamlin & Wynn (2011). Tiger puppet alternately rolls the ball to each of the bunny puppets in turn. One bunny rolls the ball back to the tiger; the other takes the ball and runs away with it.

the "box" and "ball" scenarios, when given an opportunity to reach for the prosocial or the antisocial puppet, all ages of infants robustly chose the former (Figure 20.3, last six bars).

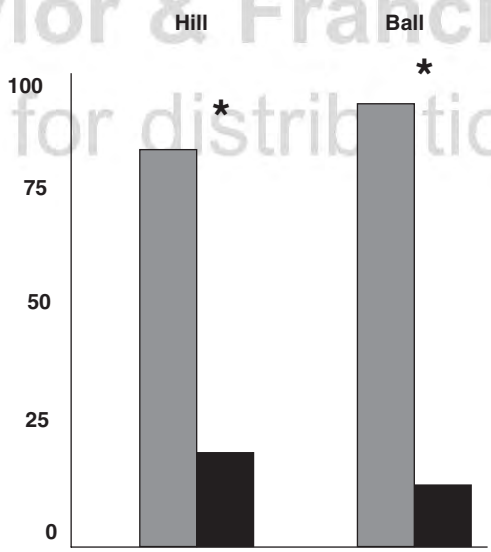
Note that the three scenarios described above differed from each other in important ways. The original "hill" scenario involved a character acting toward a location goal that was specified by its movement along a trajectory; progress along this trajectory was facilitated by the prosocial Helper, and impeded by the antisocial Hinderer, each of whom forcefully contacted the Climber with equal strength. The "boxes" scenario featured a character with a different sort of goal, that of effecting a state-change upon an object (of opening the closed box), specified by repeated-but-failed attempts to lift the lid; these same actions were repeated by the Helper, while the Hinderer responded in an entirely different manner. Finally, in the "ball" scenario, *no* goal was expressed by the protagonist: The goodness of the prosocial puppet was because it gave back a ball (which could be seen as participating in a reciprocal give-and-take); the badness of the antisocial puppet was because it ran away with the ball. Yet despite these large differences in form, detail, and structure, infants responded similarly across all three scenarios, indicating the operation of abstract, nonperceptual concepts of prosocial and antisocial behavior.

Other researchers used a similar infant-choice methodology to find that 1-year olds generate preferences after observing a very different type of social behavior—fair versus unfair allocation of resources. Geraci and Surian (2011) presented 10- and 16-month-old infants with puppet shows in which a lion and a bear distributed disks to a donkey and a cow. The lion (or bear) gave each animal one disk, and the bear (or lion) gave one animal two disks and the other nothing. The lion and the bear were then held up, and the toddler subjects were asked, "Which one is the good one? Please show me the good one." The 10-month olds guessed randomly, but the 16-month olds preferred the fair divider over the unfair one. (See also Schmidt and Sommerville, 2011, and Sloane, Baillargeon, & Premack, 2012, for looking-time studies that assess infants' and toddlers' expectations about fairness.)

What about very young babies? Just how early does the capacity to make judgments on the basis of behavior emerge? To explore this, we tested 3-month olds by showing them our "hill" and "ball" scenarios. Being so young, they were unable to reach for one of the actors to indicate their preference, and so we assessed their preference by showing them both characters simultaneously in a preferential looking paradigm and noting which one they were more drawn to visually (see Figure 20.6A). This method exploits



A



B

Figure 20.6 (A): Test display (from Hill scenario) in 3-month old preferential-looking task of Hamlin, Wynn, & Bloom (2010). Infants were shown the prosocial and antisocial characters simultaneously for 20 seconds; looking time to each was measured during this period. Procedure was the same for Ball scenario in Hamlin & Wynn (2011; test display not shown). (B): Percentage of 3-month olds who oriented more to prosocial (light bar) and to antisocial (dark bar) characters, for Hill (Hamlin, Wynn, & Bloom 2010) and Ball (Hamlin & Wynn, 2011) scenarios. Three-month olds oriented significantly toward the prosocial individual in preference to the antisocial individual, both p's < 0.05.

the fact that babies this age tend to look toward characters that they prefer, and, indeed, subsequent analyses of the 6-month olds from Hamlin, Wynn, and Bloom (2007) found that these infants, before reaching for the character, would look in the character's direction (see Hamlin, Wynn, & Bloom 2010). As predicted, we found that 3-month olds orient far more toward the prosocial actors than the antisocial ones (see Figure 20.6B)—by a factor of over two to one (Hamlin, Wynn, and Bloom, 2010; Hamlin & Wynn, 2011). Even at this young an age, then, infants' attention is already directed toward cooperative, reciprocating individuals and away from noncooperative individuals. Importantly, this is long before infants themselves have had personal experience in moving from one location to another (as the Climber is attempting in our "hill" scenario) or in reaching for and handing over objects (as the protagonist is doing in our "ball" scenario).

We also find an intriguing developmental pattern. Like 5-month-old and older infants, 3-month olds preferred a neutral character to an antisocial one; but unlike our older babies, they did *not* prefer a prosocial character to a neutral one. This suggests that judging antisocial actors as "bad" may emerge developmentally prior to assessing prosocial actors as "good": such a pattern fits well with the so-called *negativity bias* found in adults (e.g., Abelson & Kanouse, 1966; Aloise, 1993; Kanouse & Hanson, 1972; Knobe, 2003) as well as in children and infants (e.g., Hornik, Risenhoover, & Gunnar, 1987; Leslie, Knobe, & Cohen, 2006; Mumme & Fernald, 2003; Vaish, Grossmann, & Woodward, 2008), in which negative social information is more salient, more rapidly attended to, remembered more readily, and more heavily weighted than is positive information.

The Role of Intentions

The studies described above indicate that even within the first few months of life, humans can differentiate positive from negative interactions, and prefer actors who engage in the former from those who engage in the latter. We will consider now the precise nature of these judgments. Are they really *moral* ones?

As mentioned above, there is no consensus among philosophes and psychologists as to precisely what counts as "moral." But there are certain properties that are relatively uncontroversial. For one thing, moral or immoral actions are those done by *intentional agents*. It is a tragedy if an avalanche kills a person, but, unless one believes that someone set off the avalanche or it was the act of a malevolent deity, it is not a moral transgression. Rocks aren't moral beings. Further, moral actions are typically those that are done to, or directly influence, other intentional—or at least *sensate*—entities. Kicking a child is wrong; kicking a puppy is wrong; kicking a stone isn't wrong. Stones don't feel anything, and so cannot be wronged.

(Note that there are nuances to this consideration; our qualification of "typically" is to acknowledge the interesting category of actions that are judged as immoral but where nobody is harmed. For instance, many people believe that it is wrong to engage in certain harmless and consensual sexual acts, such as homosexual sex. And many believe that there can be moral wrongs without a sensate moral patient, as when defacing a flag, for instance—and these intuitions remain even when it's made perfectly clear that nobody is actually harmed [see, e.g., Haidt, 2012]. Such interesting cases fall outside the discussion here.)

The importance of intentional agents and sensate patients motivates us to look again at the infants' evaluations in the experiments described above. Infants in our studies *might* be

judging the characters' actions on the basis of their social impacts, but they might also simply be responding to other features of the interactions. Perhaps, for example, infants preferentially like someone going in an uphill direction because they find it more inspiring than the downhill equivalent; perhaps they like box-opening more than box-closing because box-opening affords the possibility of looking inside the box. There are many nonsocial reasons infants might have preferred the positive to the negative actor in our specific scenarios.

To address this question, we ran a series of control conditions to ascertain whether infants favor the character associated with a given action *only when that action is directed toward an animate agent*. In these control conditions, we showed infants events in which the recipient of the actions was not a social agent but instead an inanimate object.

In the control condition to our hill scenario, we showed babies (both 6- and 10-month olds, corresponding to the ages tested in the original "hill" scenario) one agent pushing an inert, inanimate object up the hill, and another agent pushing this same object down the hill; infants showed no preference for either agent (Hamlin, Wynn, & Bloom, 2007). The same lack of preference held when we showed these modified stimuli to 3-month olds; they looked equally to both (see Figure 20.7) (Hamlin, Wynn, and Bloom, 2010).

In a control condition to our box scenario, infants (5 and 9 months of age) repeatedly saw an inanimate "pincer" grip and partially raise the box lid; in some events, one puppet then grasped the lid and fully opened it, while in other events a different puppet jumped

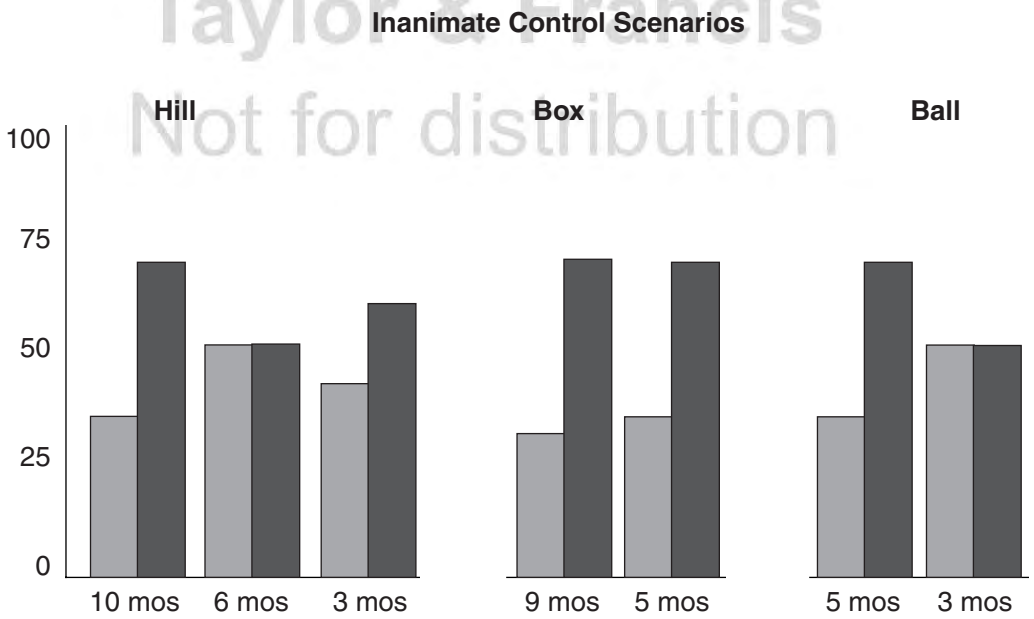


Figure 20.7 Inanimate Control Scenarios

Credit: Percentage of infants preferring the Pusher Upper/Box Opener/Ball Giver (light bars), versus the Pusher Downer/Box Closer/Ball Taker (black bars), across the three nonsocial control scenarios and for each age tested. Infants' preference patterns at all ages and for all conditions differed significantly from their preferences in the social scenarios (depicted in Figure 20.3 for 5-, 6-, 9-, and 10-month olds; and in Figure 20.6 for 3-month olds); in none of the nonsocial conditions did infants significantly prefer one character over the other.

on the partly opened lid, slamming it shut. (This control condition was based on findings showing that infants of these ages view inanimate rods and pincers very differently from intentional agents, and do not assign goals to them; see, e.g., Meltzoff, 1995; Woodward, 1998). In this modified box-opening scenario, infants no longer preferred the box opener (Hamlin & Wynn, 2011). Finally, in the inanimate control condition to our ball scenario, a "pincer" held the ball and manipulated it, dropping it so that it rolled first to one puppet and then another; these two puppets respectively placed it back into the jaws of the pincer (the "giver" puppet), and ran away with it (the "taker" puppet). Infants showed no preference, in this nonsocial context, for a ball giver over a ball taker (Hamlin & Wynn, 2011).

In sum, infants' responses in our control conditions indicate that in our experimental conditions, they were not responding to superficial perceptual aspects or physical consequences of the actions they witnessed. Rather, they were attending to the social meanings of these actions. Infants across the first year of life readily evaluate actors on the basis of how their actions influence other agents—their social effects and consequences.

One can further ask whether infants are sensitive to the *social intentions* of an actor, independent of the consequences of a social act. As adults, we distinguish intent from consequence. There is a world of difference between hitting someone by accident versus hitting someone by mistake. Piaget (1965) famously argued, though, that children much older than those we have been talking about are largely insensitive to intentions, focusing solely on consequences in their moral judgments. Is there any evidence that intentions of an actor, independent of the consequences of her action, influence infants' evaluations of the actor?

Many studies show that infants are capable of distinguishing the intent of an action from its physical consequences: They can identify the intended goal of an action whose outcome is unseen or that fails to achieve the goal (e.g., in 8-month olds: Hamlin, Newman, & Wynn, 2009; in 12- to 18-month olds: Bellagamba & Tomasello, 1999; Csibra, Biro, Koos, & Gergely, 2003; Meltzoff, 1995). But there is no direct evidence, yet, as to whether the intentions per se of an actor engaged in a third-party interaction influence infants' judgment of the actor. An appropriate experimental design to address this question could consist of showing infants one actor who is deliberately harming (or aiding) another, and another actor who clearly *accidentally* achieves the same ends, and then asking if infants prefer the clumsy-but-innocent actor to the competently malevolent one (or if they prefer the well-intentioned do-gooder to the accidental do-gooder).

Still, there is suggestive evidence that infants are sensitive to agent intent in their evaluations. In one study, infants 9 months and older (but not 6-month olds) were more impatient with an experimenter (as reflected in behaviors such looking away and disengaging) who was unwilling to give them a toy but who "teased" them with it, than with an experimenter who attempted to hand over the toy but dropped it (Behne, Carpenter, Call, & Tomasello, 2005). Another study found that infants in their second year distinguished an "unwilling" actor from an "unable" actor, being far more likely to later help the former. They were also equally willing to help (a) a "willing" experimenter who successfully managed to hand them a toy, and (b) a "willing" experimenter who tried but failed to do so (Dunfield & Kuhlmeier, 2010). That is, in this study, infants prioritized intent over outcome.

These findings indicate that infants can assess the valence of the *intention* behind an action, distinct from the valence of the action's *outcome*, and that they can generate a disposition toward an actor on the basis of his or her intent. What has not yet been investigated

is whether infants assess third parties in this manner, as opposed to just those actors with whom the infant him- or herself is engaged. But clearly both of the requisite components are present: Infants can assess the valence of third-party interactions, and they can assess the valence of intent and respond to it appropriately, abstracted away from outcome. Further research will determine when in development these components are integrated into judgments of third-party interactions.

Reward and Punishment

Our mature moral judgments of others have implications for how we think about and interact with them. One interesting implication is that we wish for good individuals to be rewarded, and feel that bad individuals are deserving of punishment.

Indeed, it has been proposed that our urge to punish or shun "evildoers"—those who violate community norms of prosociality—is an essential feature of a cooperative species, one that is required in order for reciprocal altruism and cooperation to evolve (e.g., Bowles & Gintis, 2004; Fehr & Rockenbach, 2004). Moreover, not only punishment or shunning of noncooperators themselves, but also the punishment or shunning of those who in their turn *fail to punish* noncooperators, may be an important requirement for stabilizing cooperation within a group (Henrich & Boyd, 2001).

Some have argued further that we have evolved a taste for "altruistic" or "costly" punishment, where we are willing to suffer to make another pay. This has proved to be controversial, though, as such an inclination is difficult to explain in terms of natural selection. Suppose our society works well—and to our benefit—if free riders are brought into line through punishment. But who does the punishment? Given that it's costly, we seem to have the free-rider problem all over again. What's keeping an individual from doing nothing and benefitting from the sacrifice of others, being a free rider when it comes to punishing free riders? Now, it might be that we are motivated to punish those who shirk from punishing free riders . . . but then are we also motivated to punish those who shirk from punishing those who shirk from punishing free riders? Nobody doubts that enforcement is good for the group, but it turns out to be vexingly hard to explain how it could evolve through natural selection (Dreber et al., 2008). Indeed, a recent review of the literature from sociology and anthropology finds that altruistic punishment is rare or nonexistent in the small-scale societies of the real world (Guala, 2012). There are plenty of direct and indirect ways to make wrongdoers, including free riders, suffer. But there isn't altruistic or costly punishment; real-world punishment tends to be done in ways that are not costly to the punisher, either because they don't involve confrontation (e.g., gossiping/bad-mouthing someone behind their back) or because they are imposed by the group as a whole, and so no single individual stands out as a target of potential retaliation.

Regardless of the origins of this inclination, it's clear that adults are motivated to reward do-gooders and cooperators and to punish evildoers and free riders. But when do these tendencies emerge developmentally?

For older children, we might see some hint in the behavior of tattling. Children love to tattle. In studies of siblings between the ages of 2 and 6, researchers found that most of what the children said to their parents about their brothers or sisters counted as tattling (Den Bak & Ross, 1996; Ross & Den Bak-Lammers, 1998). And their reports tended to be accurate.

Based on their study of children in an inner city school in Belfast, Ingram and Bering (2010) note that it is rare for children to talk to their teachers about something good that someone else has done; most of their reports were about negative behaviors. (This is true as well about third-person descriptions of others' behaviors by adults—gossip.) To better understand children's motivations for reporting the bad acts of their peers and siblings, it is instructive to look at the conditions under which children tattle. In one study, 2- and 3-year olds were taught a new game to play with a puppet; when the puppet started to break the rules, the children would spontaneously complain to adults (Rakoczy, Warneken, & Tomasello, 2008). Another study found that 3-year olds tended to tattle when someone destroyed an artwork that someone else made, but not when the individual destroyed an artwork that nobody cared about (Vaish, Missana, & Tomasello, 2011).

Still, the phenomenon of tattling does not prove that children are interested in just punishment; among other things, they might want to show themselves off to adults as good moral agents, as responsible beings who are themselves sensitive to right and wrong. (It is an interesting question whether children would tattle *anonymously*.) And these children are relatively old compared to the research we've been discussing so far. Is there a way to assess whether a desire exists, in infants, to see just punishment meted out?

To address this question, and to ask whether infants are assessing not only the "pleasantness" of one individual's act toward another but also the *rightness* or *goodness* of it, we have carried out a series of studies investigating infants' and children's inclinations to reward and punish, as well as their assessment of *others* who reward or punish appropriately, compared to those who reward or punish inappropriately (Hamlin et al., 2011).

In one study, we found that the social behavior of an individual toward a third party influences how toddlers wish to treat that individual. After observing interactions in which one puppet acts prosocially and another acts antisocially, 21-month-old toddlers were presented with the opportunity to give a single treat to (i.e., reward) the character of their choice. They robustly chose the more positive character. But when asked to *take* a treat from (i.e., punish) one of the individuals, toddlers chose the more negative character. The desires to treat well those who do good unto others, and to punish those who violate our social norms of cooperation, are already operative in young childhood.

In a further set of studies, we asked how both toddlers and infants want *others* to treat prosocial and antisocial individuals. We first showed subjects—infants 5 and 8 months of age—scenarios in which a puppet was trying to open a box; on some occasions a Helpful puppet joined in and helped get the box open, and on other occasions a Hindering puppet jumped on the box lid, slamming it shut. Would subjects wish to have the Helper treated nicely (rewarded), but to have the Hinderer treated badly (punished)? To ask this, we next showed subjects *either* the Helper, *or* the Hinderer, in a situation in which it was (in its own turn) treated positively by one new character and negatively by another. Which of these characters would the subject prefer? The experimental events unfolded as follows: After showing its initial prosocial or antisocial tendencies in the "box" scenario, either the Helper or the Hinderer became an actor in a *second* scenario, in which it played with a ball. It rolled the ball, on alternate trials, toward a new puppet who returned the ball (a "Rewarding puppet"), and toward a different new puppet who kept the ball and ran away with it (a "Punishing puppet"). Subjects were then allowed to choose between the Rewarding and Punishing puppets. Would they choose the Rewarder over the Punisher, when these

puppets were acting on the Helper from the previous scenario? (This might be expected in any case, given that infants appear so robustly to prefer prosocial behaviors). And, of great interest, would they prefer the Punisher over the Rewarder when the two puppets were acting on the Hinderer—a character previously observed to have been antisocial?

It's possible that infants would prefer the Rewarder in both situations, if they were sensitive only to the "local" valence of a social interaction. Giving the ball back is a locally positive act, while taking the ball and running off with it is locally negative. Infants may assess the local valence of an action and respond solely on that basis, without regard to the larger context in which it occurs. Indeed, this is precisely what the 5-month olds did; they attended only to the local valence, and preferred the Rewarder regardless of who it was rewarding.

But the 8-month olds were more sophisticated. They preferred someone who was nice to a prosocial individual over one who was mean (75% chose the Rewarder of the prosocial character); but they also preferred an individual who was mean to an antisocial individual, over one who was nice (81% chose the Punisher of the antisocial character). Even within the first year of life, then, the valence of an individuals' social behavior influences how we want him or her to be treated, and influences how we judge *others* who treat that individual positively or negatively.

The finding that even infants 8 months of age prefer individuals who treat do-gooders well, but prefer those who treat evildoers badly highlights an important aspect of our social evaluative judgments. We do not judge an individual solely on the basis of the positive or negative *local* value of his or her action, taken in isolation. It is the broader context in which a prosocial or antisocial action occurs that determines its ultimate valence. We—even as infants—*do* favor negative social actions, when the target of the action is someone who has themselves behaved in an antisocial manner.

Final Word: But Are These Really *Moral* Judgments?

Babies show concern at others' pain and sorrow, make spontaneous efforts to console others, and spontaneously help others even at external costs to themselves, suggesting that helping others is intrinsically rewarding. Within the first few months of life, infants can identify whether a social interaction that takes place between third parties is a positive or a negative one, and this influences their own attitude (attraction or aversion) to whoever initiated the interaction. It also influences their attitude toward other individuals who engage with the actor—what we might call infants' second-order social judgments. Infants are drawn to those who do good for others, and have an aversion to those who do bad to others; they themselves (at least by toddlerhood ages) prefer to bestow negative treatment upon one who has acted badly toward another, and to bestow positive treatment upon one who has acted prosocially; and, before their first birthdays, they prefer *others* who reward the good and who punish the bad.

In all these ways, infants' and toddlers' social judgments and responses bear a strong resemblance to those of adults. The early emergence of the evaluation of social actions—present already by 3 months of age—suggests that this capacity cannot result entirely from experience in particular cultural environments or exposure to specific linguistic practices, and it suggests that there are innate bases that ground some components of our moral cognition.

Still, babies have a long way to go. The sentiments and evaluations we have reviewed do not comprise a full-fledged system of moral reasoning. We often care about the fate of distant strangers, something which is likely not present in babies or young children, and likely not part of our evolved nature (see Bloom, in press). As David Hume put it, "Were we, therefore, to follow the natural course of our passions and inclinations, we should perform but few actions for the advantage of others . . . because we are naturally very limited in our kindness and affection (Hume, 1739/2000, p. 519).

Also, under many analyses, adult morality can encompass concerns about domains such as sexual purity and religious sanctity (e.g., Bloom, in press; Haidt, 2012; Shweder et al., 1997). Acts such as consensual sexual relations between people of the same sex evoke the same sort of reactions as more prototypical moral violations, such as one person hitting another—outrage, anger, and a desire that the perpetrator be punished. Although we have never looked, we think it unlikely that these reactions will show up early in development.

Finally, adults can reason according to impartial codes of fairness and justice, and we can consciously develop systems that dictate appropriate moral action, as we find in law, religion, and philosophy, and, again, we don't find this in babies or young children.

Still, what we *do* find—the capacity to evaluate an individual's social action as positive or negative, and to generate attitudes toward others based on these evaluations—comprises an essential basis of any moral system that will eventually contain more abstract concepts of right and wrong. We would never be moral beings if we did not start as moral babies.

References

Abelson, R.P. & Kanouse, D.M. (1966). Subjective acceptance of verb generalizations. In S. Feldman (Ed.), *Cognitive consistency* (p. 173–199). San Diego, CA: Academic Press.

Aloise, P.A. (1993). Trait confirmation and disconfirmation: The development of attribution biases. *Journal of Experimental Child Psychology, 55,* 177–193.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Baillargeon, R. (1987). Object permanence in 3 1/2 and 4 1/2 month old infants. *Developmental Psychology, 23,* 655–664.

Baron-Cohen, S. (2004). *The essential difference: Male and female brains and the truth about autism*. New York: Basic Books.

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology, 41*(2), 328–337.

Bellagamba, F., & Tomasello, M. (1999) Re-enacting intended acts: Comparing 12- and 18 month-olds. *Infant Behavior and Development, 22*(2), 277–282.

Bloom, P. (in press). *Just babies: The origin of good and evil*. New York: Crown.

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology, 65,* 17–28.

Csibra, G., Biro, S., Koos, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science, 27,* 111–133.

Darwin, C. (1877). A biographical sketch of an infant. *Mind, 2,* 285–294.

Den Bak, I. M., & Ross, H. S. (1996). "I'm telling!" The content, context, and consequences of children's tattling on their siblings. *Social Development, 5,* 292–309.

Dondi, M., Simion, F., & Caltran, G. (1999). Can newborns discriminate between their own cry and the cry of another newborn infant? *Developmental Psychology, 35,* 418–426.

Dreber, A., Rand, D.G., Fudenberg, D., & Nowak M.A. (2008) Winners don't punish. *Nature, 452,* 348–351.

Dunfield, K.A., & Kuhlmeier, V.A. (2010). Intention-mediated selective helping in infancy. *Psychological Science, 21,* 523–527.
Fehr, E., & Rockenbach, B. (2004). Human altruism: Economic, neural, and evolutionary perspectives. *Current Opinion in Neurobiology, 14,* 784–790.
Geraci A., & Surian L. (2011). The developmental roots of fairness: Infants' reactions to equal and unequal distributions of resources. *Developmental Science, 14,* 1012–1020.
Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56,* 165–193.
Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences, 35,* 1–15.
Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York: Pantheon Books.
Hamlin, J.K., Newman, G., & Wynn, K. (2009). Eight-month-old infants infer unfulfilled goals, despite contrary physical evidence. *Infancy, 14,* 579–590.
Hamlin, J.K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development, 26,* 30–39.
Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450,* 557–560.
Hamlin, J.K., Wynn, K., & Bloom, P. (2010). Three-month-old infants show a negativity bias in social evaluation. *Developmental Science, 13,* 923–929.
Hamlin, J.K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences, 108,* 19931–19936.
Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208,* 79–89.
Hoffman, M. L. (2000). *Empathy and moral development*. New York: Cambridge University Press.
Hornik, R., Risenhoover, N., & Gunnar, M. (1987). The effects of maternal positive, neutral, and negative affective communications on infant responses to new toys. *Child Development, 58,* 937–944.
Hume, D. (1739/2000). *A treatise of human nature*. New York: Oxford University Press.
Ingram, G. P. D., & Bering, J. M. (2010). Children's tattling: The reporting of everyday norm violations in preschool settings. *Child Development, 81,* 945–957.
Kanouse, D.M., & Hanson, L.R. (1972). Negativity in evaluations. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 47– 62). Morristown, NJ: General Learning Press.
Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63,* 190–193.
Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-old infants. *Psychological Science, 14,* 402–408.
Kuhlmeier, V., Wynn, K., & Bloom, P. (under review). Nine- and 12-month-olds interpret goal-directed actions based on past interactions.
Leslie, A., Knobe, J. & Cohen, A. (2006). Acting intentionally and the side-effect effect: "Theory of mind" and moral judgment. *Psychological Science, 17,* 421–427.
Martin, G., & Clark, R. (1982). Distress crying in neonates: Species and peer specificity. *Developmental Psychology, 18,* 3–9.
Masserman, J. H., Wechkin, S., & Terris, W. (1964). "Altruistic" behavior in rhesus monkeys. *American Journal of Psychiatry, 121,* 584–585.
Meltzoff, A.N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31*(5), 838–850.
Mumme, D., & Fernald, A. (2003). The infant as onlooker: Learning from emotional reactions observed in a television scenario. *Child Development, 74,* 221–237.
Nado, J., Kelly, D., & Stich, S. (2009). Moral judgment. In J. Symons & P. Calvo (Eds.), *The Routledge companion to the philosophy of psychology* (pp. 621–633). New York: Routledge.
Nowak, M.A. & Highfield, R. (2011). *SuperCooperators: Why we need each other to succeed*. New York: Simon & Schuster.
Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308,* 255–258.

Piaget, J. (1965). *The moral judgment of the child*. New York: The Free Press.
Premack, D., & Premack, A.J. (1997). Infants attribute value +/– to the goal-directed actions of self-propelled objects. *Journal of Cognitive Neuroscience, 9,* 848–856.
Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology, 44,* 875–881.
Rice, G. E. (1964). Aiding behavior vs. fear in the albino rat. *Psychological Record, 14,* 165–170.
Rice, G. E., & Gainer, P. (1962). "Altruism" in the albino rat. *Journal of Comparative and Physiological Psychology, 55,* 123–125.
Ross, H. S., & Den Bak-Lammers, I. M. (1998). Consistency and change in children's tattling on their siblings: Children's perspectives on the moral rules and procedures of family life. *Social Development, 7,* 275–300.
Sagi, A., & Hoffman, M. (1976). Empathic distress in the newborn. *Developmental Psychology, 12,* 175–176.
Schmidt, M., & Sommerville, J. (2011). Fairness expectations and altruistic sharing in 15-month-old human infants. *PLoS ONE 6*(10), e23223.
Shweder, R., Much, N., Mahapatra, N., & Park, L. (1997). The "big three" of morality (autonomy, community, and divinity), and the "big three" explanations of suffering, as well. In A. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). New York: Routledge.
Sinnott-Armstrong, W., & Wheatley, T. (in press). The disunity of morality and why it matters to philosophy. *Monist*.
Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness? *Psychological Science, 23,* 196–204.
Smith, A. (2002). *The theory of moral sentiments.* New York: Cambridge University Press. (Original work published 1759)
Spelke, E. (1990). Principles of object perception. *Cognitive Science, 14,* 29–56.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46,* 35–57.
Trivers, R.L. (1985). *Social evolution*. Reading, MA: Benjamin/Cummings.
Turiel, E. (1998). The development of morality. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development*. New York: Wiley.
Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin, 134,* 383–403.
Vaish, A., Missana, M., & Tomasello, M. (2011). Three-year-old children intervene in third-party moral transgressions. *Developmental Psychology, 29,* 124–130.
vanMarle, K., & Wynn, K. (2006). Six-month-old infants use analog magnitudes to represent duration. *Developmental Science, 9,* F41–F49.
Warneken, F., & Tomasello, M. (2006). Altrustic helping in human infants and young chimpanzees. *Science, 311,* 1301–1303.
Warneken, F., & Tomasello, M. (2008). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. *Developmental Psychology, 44,* 1785–1788.
Warneken, F., & Tomasello, M. (2009). The roots of human altruism. *British Journal of Psychology, 100,* 455–471.
Wechkin, S., Masserman, J. H., & Terris, W., Jr. (1964). Shock to a conspecific as an aversive stimulus. *Psychonomic Science, 1,* 47–48.
Woodward, A. (1998). Infants selectively encode the goal of an actor's reach. *Cognition, 69,* 1–34.
Wynn, K. (1992). Addition and subtraction by human infants. *Nature, 358,* 749–750.
Wynn, K. (2008). Some innate foundations of social and moral cognition. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Foundations and the future* (pp. 330–347). Oxford: Oxford University Press.
Yamaguchi, M., Kuhlmeier, V., Wynn, K., & vanMarle, K. (2009). Continuity in social cognition from infancy to childhood. *Developmental Science, 12,* 746–752.
Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992a). Development of concern for others. *Developmental Psychology, 28,* 126–136.
Zahn-Waxler, C., Robinson, J. L., & Emde, R. N. (1992b). The development of empathy in twins. *Developmental Psychology,* 28, 1038–1047.